Jerry A. Fodor

El lenguaje del pensamiento

Hito bibliográfico de la Psicología Cognitiva, la publicación —en 1976— de esta obra significó uno de los primeros intentos serios para sentar las bases teóricas de una psicología no conductista. Su tesis central sigue vigente y ha contribuido en buena medida al desarrollo de la investigación durante los últimos años. El hilo argumental que recorre todo el libro es la afirmación según la cual el punto de vista computacional es el único modo plausible de entender la actividad mental de los organismos. En una extensa introducción, Jerry A. Fodor establece el nivel propio de teorización en psicología, contraponiéndolo al reduccionismo conductista y al fisicista, y defiende la necesidad de un lenguaje mentalista. El capítulo primero examina tres ejemplos de actividad mental (la acción deliberada, el aprendizaje de conceptos y la percepción) cuyo estudio exige la postulación de procesos computacionales dotados de un sistema representacional o código interno. A continuación, la obra analiza la naturaleza y propiedades de este sistema, distinguiendo entre el lenguaje del pensamiento y el lenguaje propiamente dicho. Los dos últimos capítulos recapitulan la evidencia empírica procedente de la lingüística generativa y de la psicología cognitiva, a fin de comprobar hasta qué punto los datos apoyan la teoría de las representaciones internas. Una conclusión final reflexiona sobre el alcance del enfoque computacional; aunque este modo de hacer psicología ofrece limitaciones, no parece existir otra alternativa viable a la vista. El interés del libro, cuyo estilo desenfadado y casi coloquial es compatible con el rigor, desborda el ámbito de la psicología y penetra en el campo de la lingüística, la filosofía, las ciencias de la computación y las ciencias neurológicas.

Alianza *Psicología*

El lenguaje del pensamiento

Alianza Psicología

Jerry A. Fodor

El lenguaje del pensamiento

Versión española de Jesús Fernández Zulaica
Presentación y revisión técnica de José E. García-Albea

Alianza
Editorial

Título original: *The Language of Thought.* Esta obra se publica por acuerdo con Harper & Row, Publishers, Inc., de Nueva York.

Calle Milán, 38; ☎ 200 00 45
ISBN: 84-206-6506-1
Depósito legal: M. 4.544-1985
Fotocomposición: EFCA
Impreso en Hijos de E. Minuesa, S. L.
Ronda de Toledo, 24. 28005 Madrid
Printed in Spain

Sólo conectar
E. M. FORSTER

Está cambiando el viento
MARY POPPINS

Gran parte de lo que aparece en el presente libro es fruto de las conversaciones mantenidas en ratos sueltos (y en contextos muy dispares) con mi esposa, Janet Dean Fodor. A ella va dedicado con amor y gratitud.

INDICE

Presentación, por José E. García-Albea 11
Prefacio 15
Prefacio a la edición castellana 19
Introduccióu: Dos clases de reduccionismo 23

Conductismo lógico, 24. Reduccionismo fisiológico, 31.

1. Primeras aproximaciones 47
2. Lengnaje privado, lenguajes pńblicos 73

Por qué tiene que haber un lenguaje privado, 73. Cómo podría darse un lenguaje privado, 82. Cómo debe ser el lenguaje privado, 96.

3. La estruetnra del código interno: Algunas pruebas lingüísticas 115

El vocabulario de las representaciones internas, 138.

4. La estructura del código interno: Algunas pruebas psicológicas 173
Conclusión: Alcance y límites 213
Bibliografía 221
Indice alfabético 227

PRESENTACION

Tras no pocas vicisitudes, es una gran satisfacción poder ofrecer, al lector de habla hispana, una de las obras claves para entender por dónde va y a dónde se dirige la psicología científica actual. EL LENGUAJE DEL PENSAMIENTO fue publicado en su versión original hace casi diez años y puede ser considerado como uno de los hitos bibliográficos de lo que se ha dado en llamar «Psicología Cognitiva». Y ello, por dos razones principales. En primer lugar, porque, en su momento, supuso quizá el primer intento serio de profundizar en las bases teóricas de la psicología que se venía haciendo, desde algún tiempo antes, como alternativa al enfoque conductista. Y en segundo lugar, porque, visto desde ahora, se puede constatar que el argumento central del libro continúa vigente y ha contribuido, en buena medida, al desarrollo de la investigación psicológica que se ha producido en los últimos años.

En síntesis, la obra de Fodor hace explícitos los presupuestos del paradigma cognitivo y trata de extraer de ellos sus implicaciones últimas, sin perder contacto en ningún momento con los principales resultados empíricos disponibles. La idea maestra que recorre de principio a fin todo el libro, y en la que vuelve a insistir Fodor en su prólogo a esta edición española, se podría formular así: No hay modo plausible de entender la actividad mental de los organismos si no es desde un punto de vista computacional, es decir, como conjunto de operaciones formales que versan sobre símbolos o representaciones.

El libro está compuesto de una extensa introducción, cuatro largos capítulos y una conclusión final. En la introducción, Fodor trata de situar el nivel propio de teorización en psicología, saliendo al paso de los dos tipos de reduccionismo que han amenazado con frecuencia la integridad de esta ciencia, el reduccionismo conductista y el reduccionismo fisicista. Frente a ellos, el autor defiende la necesidad de un lenguaje mentalista en el que plasmar el papel determinante que le corresponde al sujeto en la causación de su conducta, a través de sus estados y procesos internos. En los cuatro capítulos que constituyen el núcleo de EL LENGUAJE DEL PENSAMIEN-

TO, la tarea que se impone Fodor va a ser precisamente la de sentar las bases para una teoría de esa vida mental que subyace al comportamiento manifiesto de los organismos.

En el capítulo primero, Fodor se detiene a considerar tres ejemplos de actividad mental (la acción deliberada, al aprendizaje de conceptos y la percepción), en los que se puede comprobar que las teorías disponibles no tienen más remedio que postular procesos computacionales que, como tales, requieran un medio en que poderse llevar a cabo, es decir, un sistema representacional. Este sistema representacional o código interno es lo que Fodor denomina «el lenguaje del pensamiento».

En el segundo capítulo, se va a tratar de la naturaleza y propiedades generales de dicho sistema representacional. Para ello, Fodor empieza por distinguir claramente entre el lenguaje del pensamiento y el lenguaje propiamente dicho (es decir, las lenguas naturales). Aparte de los argumentos clásicos para mantener la distinción entre pensamiento y lenguaje, el autor va a recurrir precisamente al hecho de la adquisición del lenguaje natural para demostrar la necesidad de postular un lenguaje del pensamiento. Y ello, en parte, por las mismas razones utilizadas en el capítulo anterior con respecto a los tres casos allí examinados. Entre las propiedades fundamentales de ese lenguaje interno, destacan su carácter innato y su elevado poder resolutivo. Para entender mejor el sentido de estas propiedades, Fodor recurre a la famosa metáfora del computador e insiste en la necesidad de interpretarla literalmente. El lenguaje del pensamiento vendría dado con el organismo, del mismo modo que todo computador está construido con su propio lenguaje-máquina. Este es el lenguaje en el que se efectúan las operaciones internas del sistema y que posibilita, mediante algún mecanismo de traducción (al modo de los compiladores del computador), la comprensión y adquisición del lenguaje natural (en el caso de los computadores, la utilización de los lenguajes de programación).

Los capítulos tercero y cuarto van a estar dedicados a recapitular una buena parte de la evidencia empírica procedente de la lingüística generativa y de la psicología cognitiva, para ver hasta qué punto los datos disponibles apoyan la teoría de las representaciones internas y en qué medida constriñen aún más dicha teoría.

Por lo que respecta a los datos lingüísticos (cap. 3), la estrategia que va a seguir Fodor se ajusta, en principio, a la más ortodoxa tradición de considerar el estudio del lenguaje como la gran avenida que nos lleva a estudiar el pensamiento. Y precisamente lo que va a tratar de ver es cómo los datos sobre el lenguaje nos revelan algunas de las propiedades del pensamiento; en concreto, que el pensamiento es en sí mismo un lenguaje —con su propia sintaxis y con su propio vocabulario—, independiente y no reducible al lenguaje natural. Los hechos lingüísticos que analiza Fodor están relacionados con el tema de la composicionalidad de los conceptos, defendiendo la postura de que el sistema representacional interno habría que entenderlo en términos similares a lo que Carnap llamaba «postulados del significado» (reglas de inferencia sobre representaciones), más que entenderlo como conjunto de definiciones. La consecuencia más directa de todo ello es la de ampliar enormemente la base primitiva del vocabulario interno, aumentando así su capacidad representacional.

Por lo que respecta a los datos psicológicos (cap. 4), Fodor va a efectuar un amplio repaso de la investigación experimental realizada en diversas áreas de la psicología cognitiva, con el fin de mostrar su incidencia sobre una teoría de las representa-

ciones mentales. La estrategia, en este caso, no es la de ir buscando con lupa tal o cual resultado concreto que le confirme alguna particularidad de su teoría, sino, más bien, la de enfrentarse de lleno con resultados provocativos, y a menudo contradictorios, que ponen claramente de manifiesto, no obstante, una de las propiedades generales más características del sistema representacional, a saber, la de su gran flexibilidad. Flexibilidad que va a estar caracterizada, a su vez, por la *racionalidad* con la que el organismo explota los recursos que le proporciona su sistema representacional, es decir, por el uso inteligente que hace de las representaciones internas. Fodor aprovecha una buena parte de este capítulo para entrar en la discusión acerca del formato de representación, examinando asimismo los resultados más relevantes de la investigación sobre las imágenes mentales. Su argumentación le llevará finalmente a resaltar la necesidad de apelar a un código *común,* de naturaleza *proposicional, intermodal* y suficientemente *abstracto,* considerando a éstos como atributos característicos del lenguaje del pensamiento.

En su conclusión final, Fodor hace una reflexión sobre el ámbito y los límites de la teorización psicológica, tal y como se desprenden del examen al que ha sometido a la psicología cognitiva. Por una parte, esta nueva forma de entender la psicología estaría en conexión bastante directa con el punto de vista tradicional (e incluso con el punto de vista vulgar) sobre el objeto propio de esta ciencia: la vida mental. Pero por otra parte, su manera de abordar dicho objeto y los medios con los que ahora ya cuenta para estudiarlo marcan un cambio sustancial con respecto a los antiguos intentos: la vida mental se entiende ahora como actividad computacional en sentido estricto. Es indudable que ello puede suponer una constricción seria a las posibles aspiraciones de una psicología que se resista a aceptar que el organismo quede reducido a un sistema computacional. Fodor es consciente de esta constricción y es el primero en hacer explícitas las limitaciones con que se encuentra esta manera de hacer psicología. Sin embargo, y esto es lo que pretende comunicar a lo largo de todo el libro, no parece que haya otra alternativa viable en perspectiva.

No cabe duda de que EL LENGUAJE DEL PENSAMIENTO es un libro que ha de suscitar la discusión y la crítica. Ahora bien, conviene advertir que su lectura no es fácil y requiere cierta concentración para seguir los distintos argumentos y estar en disposición de contestarlos. Aun con ello, el libro resulta enormemente atractivo y lleno de sugerencias, ya que, sin sacrificar ni un punto de rigor, el estilo es desenfadado y muy próximo, en ocasiones, al lenguaje coloquial. Es de destacar, sobre todo, la forma clara y directa con que Fodor plantea las principales cuestiones que afectan al sentido mismo de la psicología cognitiva (y quizá al sentido mismo de toda psicología).

Por otra parte, es oportuno mencionar el carácter interdisciplinar de esta obra. Creemos que su interés rebasa el ámbito de la psicología y que puede alcanzar a disciplinas relacionadas, como la lingüística y la filosofía, de modo más inmediato, o como las ciencias de la computación y las ciencias neurológicas, de forma algo más indirecta. Ello se corresponde con la amplia y sólida preparación del autor en estos campos. Jerry A. Fodor es actualmente profesor del Instituto Tecnológico de Massachusetts (M.I.T.) y pertenece tanto al departamento de Psicología como al de Lingüística y Filosofía; en el primero de ellos, es a su vez director del laboratorio de Psicolingüística. Durante los últimos veinte años, Fodor ha venido publicando asidua-

mente en las principales revistas de estas disciplinas. Por lo que respecta a sus libros, además de EL LENGUAJE DEL PENSAMIENTO hay que destacar otros títulos como «Psychological Explanation» (1968), traducido ya al español, «The Psychology of Language» (1974, en colaboración con T. Bever y M. Garrett), «Representations» (1980) y «The Modularity of Mind» (1983).

JOSE E. GARCIA-ALBEA
Universidad Complutense de Madrid

PREFACIO

En tiempos pasados había una disciplina denominada psicología especulativa. No se podía decir del todo que fuera filosofía, ya que trataba de la construcción de teorías empíricas. Tampoco era del todo psicología, dado que no era una ciencia experimental. Sin embargo utilizaba los métodos de la filosofía y de la psicología, pues se mantenía fiel a la idea de que las teorías científicas deben tener al mismo tiempo disciplina conceptual y rigor empírico. Lo que hacían los psicólogos especulativos era lo siguiente: Pensaban en los datos existentes sobre los procesos mentales, y pensaban en las teorías psicológicas de primer orden propuestas para explicar tales datos. Luego intentaban poner en claro la concepción general de la mente que aparecía implícita en los datos y en las teorías. En general, la psicología especulativa era algo muy positivo: William James y John Dewey fueron psicólogos especulativos, lo mismo que, en algunas ocasiones, Clark Hull. Pero hoy día se suele decir que ya no quedan psicólogos especulativos.

En la medida en que es cierto que no los hay, el hecho es fácil de explicar. En primer lugar, la psicología especulativa manifestaba la inestabilidad consustancial a un híbrido. La distinción entre teorías de primer orden y teorías de orden superior es en gran parte heurística en toda ciencia no formalizada, por lo que la psicología especulativa tendía a confundirse con la psicología normal y corriente. La dilucidación de los conceptos generales es una preocupación típicamente filosófica, por lo que la psicología especulativa tendía a confundirse con la filosofía de la mente. En consecuencia, los psicólogos especulativos tenían dificultades para determinar en qué departamento debían estar, con gran desconcierto de los decanos.

Además, se habían puesto de moda ciertas teorías epistemológicas —teorías sobre la forma adecuada de orientar la ciencia— que indicaban que ninguna investigación respetable *podía* ser en parte conceptual y en parte empírica, tal como se suponía que era la psicología especulativa. Según estas teorías, las cuestiones de hecho son, por principio, diferentes de las relaciones entre las ideas, y por consiguiente deberían

considerarse por separado en la práctica científica. Los filósofos que aceptaban esta epistemología podían acusar a los psicólogos especulativos de psicologizar, y los psicólogos que la aceptaban podrían acusarles de filosofar. Como, según los epistemólogos, las actividades de psicologizar y filosofar son mutuamente incompatibles, estas acusaciones produjeron un gran malestar.

Dicho en pocas palabras, hubo una época en que la psicología especulativa pasaba por ser una anomalía metodológica y un engorro administrativo. A pesar de ello, la tradición especulativa no desapreció completamente ni en la psicología ni en la filosofía de la mente. Los psicólogos empíricos siguieron realizando sus experimentos e interpretando sus datos a la luz de una concepción determinada, aunque vaga, de cómo es la mente. (Estas concepciones solían explicitarse en el curso de las discusiones metodológicas, a las que los psicólogos son muy dados.) De la misma manera, aunque hay filósofos que se declaran partidarios del análisis puro, hay otros filósofos que no. En cuanto a estos últimos, se considera que la consonancia general con los hechos relacionados con los estados mentales constituye un condicionamiento sobre las teorías de la lógica de las atribuciones de estados mentales. E incluso los filósofos analíticos leen de vez en cuando las obras de carácter empírico y dictaminan lo que significan los datos. En realidad, son muchas veces los psicólogos que se confiesan ateóricos los que luego aparecen con las pretensiones filosóficas más grandiosas (véase, por ejemplo, Skinner en *Beyond Freedom and Dignity*, 1971), lo mismo que son los aprioristas metodológicos en filosofía los que suelen mantener puntos de vista más rígidos sobre la forma de interpretar los datos (véase, por ejemplo, Malcom en *Dreaming*, 1962).

En cualquier caso, este libro es, sin disimulos de ninguna clase, un ensayo de psicología especulativa. Más en concreto, es un intento de decir cómo funciona la mente en la medida en que en los últimos estudios empíricos del lenguaje y la cognición aparecen respuestas a esa pregunta. Me parece que vale la pena intentarlo por dos razones: en primer lugar, porque la pregunta de cómo funciona la mente es enormemente interesante, y la mejor psicología es, ipso facto, la mejor respuesta que podemos ofrecer. En segundo lugar, porque la mejor psicología con que contamos es todavía una investigación en marcha, y yo estoy interesado en el progreso de dicha investigación.

Durante los diez últimos años, más o menos, hemos asistido a una proliferación de investigaciones psicológicas basadas en la opinión de que muchos procesos mentales son procesos computacionales, por lo que gran parte de la «conducta cognitiva superior» está dirigida por reglas. Las técnicas de análisis de las conductas sometidas a reglas son ahora muy conocidas entre científicos pertenecientes a disciplinas diferentes: lingüística, psicología de la simulación, psicología cognitiva, psicolingüística, etc. Lo único que se puede decir es que el empleo de estas técnicas ha revolucionado la práctica y la teoría de las ciencias de la conducta. Pero aunque es fácil ver que las cosas han cambiado, no lo es tanto decir en qué sentido lo han hecho. Mi impresión particular es que muchos profesionales tienen cada vez menos claro el carácter general del marco de referencia teórico en que se mueven y están muy poco seguros de lo que va a ocurrir más adelante. Por ello, no parece un despropósito tratar de conseguir cierta consolidación.

Eso es, supongo yo, una de las cosas para las que sirve la psicología especulativa.

Se trata de conseguir una visión lo más clara posible de la dirección de las investigaciones actuales para así tener una guía en los futuros estudios. Naturalmente, esto es muy distinto de un simple *resumen* de las investigaciones. Lo que queremos decir es: «Si nuestra psicología, en términos generales, es exacta, en ese caso la naturaleza de la mente ha de ser, más o menos, ésta: ...» y luego rellenar el espacio en blanco. Dada la aclaración especulativa, el experimentalista puede actuar en sentido contrario: «Si la naturaleza de la mente es aproximadamente..., en ese caso nuestra psicología debería ser, más o menos, así: ...», donde este último espacio en blanco habría que completarlo con nuevas teorías de primer orden. En la ciencia, los progresos se hacen gracias a los esfuerzos realizados desde distintos campos y con resultados mutuamente beneficiosos.

Concebida de esta manera, la psicología especulativa no tiene nada de infalible. En primer lugar, como busca, fundamentalmente, hacer una extrapolación de las teorías científicas existentes, corre el riesgo de que estas teorías resulten falsas. Después de todo, podría ocurrir que todo el enfoque del procesamiento de la información aplicado a la psicología fuera quizá una idea poco feliz. En ese caso, es altamente improbable que sean verdaderas las teorías de la mente inspiradas en dicho enfoque. Ya ha ocurrido esto alguna vez en la psicología. Parece ahora razonablemente claro que el enfoque de las teorías del aprendizaje aplicado a la explicación de la conducta fue una idea desafortunada, y que la teoría de la mente que proponía era grotesca. No hay nada que hacer al respecto, como no sea continuar adelante y tratar de encontrar algo mejor. Quizá la mejor forma de demostrar que nuestra psicología anda descaminada, si es que eso fuera cierto, es proponer de forma explícita la forma de entender la mente que lleva consigo.

En segundo lugar, es seguro que hay más de una manera de interpretar la moraleja de las recientes investigaciones psicológicas. En este libro voy a esbozar una teoría sobre los procesos mentales, y argumcntaré diciendo que tal teoría *está* implícita en las afirmaciones aceptadas hoy en día por la mayoría de los psicólogos cognitivos o de los lingüistas sensatos, aunque es muy posible que no todos los psicólogos cognitivos o lingüistas sensatos estén de acuerdo. En realidad, mi mayor esperanza es que este libro sirva para provocar la discusión sobre estos puntos. Algunas de las cosas que parecemos aceptar me parecen, hablando con franqueza, un tanto extrañas. Me encantaría saber si hay algún procedimiento que permita salvar la psicología sin tener que aceptar esas ideas extrañas.

Finalmente, en cuanto psicólogo especulativo, lo que uno busca es elaborar teorías empíricas de la mente que sean filosóficamente respetables, aunque quizá no puedan evitar cierta tendenciosidad filosófica. Pero, evidentemente, hay más de una manera de considerar qué es lo que constituye la respetabilidad filosófica, y hay que optar por una de ellas. En el presente libro mi argumentación se hace partiendo del presupuesto de que el realismo es mejor filosofía de la ciencia que el reduccionismo, y que, en general, no es conveniente que los filósofos intenten hacer afirmaciones ontológicas sirviéndose de argumentos epistemológicos. Reconozco, sin embargo, la (mera) posibilidad de que esta presuposición sea errónea. Si lo es, la explicación de los hechos mentales que voy a proponer estará completamente descaminada.

Este libro no es del todo culpa mía. En primer lugar, es en gran parte continua-

ción de un libro que escribí con los profesores T. Bever y M. Garrett (Fodor, Bever y Garrett, 1974). Muchas de las ideas examinadas aquí surgieron gracias a la experiencia de escribir, con ellos, un amplio trabajo de revisión de las actuales publicaciones de carácter experimental y teórico en el campo de la psicolingüística. A lo largo de nuestro trabajo, volvíamos a tratar una y otra vez de los fundamentos de la disciplina. Aquí hemos reproducido gran parte del contenido de aquellas conversaciones. Es tanto lo que debo a mis coautores que la presente obra casi podría considerarse un plagio, y mi gratitud por lo que me enseñaron es ilimitada.

Aun así, este libro no habría llegado a escribirse si no hubiera sido por una beca sabática concedida por el Massachusetts Institute of Technology (M.I.T.) y una subvención concurrente de la Fundación Guggenheim, que me permitieron liberarme de mis obligaciones académicas durante el año 1973-1974. Por eso quiero expresar aquí cumplidamente mi agradecimiento a ambas instituciones.

He hecho pruebas con algunas versiones iniciales de parte del material contenido en este libro en una serie de conferencias impartidas en el Departamento de Psicología de la Universidad de Oxford y en el Departamento de Filosofía del University College, de la Universidad de Londres. Me gustaría manifestar mi agradecimiento a la doctora Ann Treisman por haber organizado las primeras conferencias y al profesor Richard Wollheim por las segundas; también a los alumnos y a la facultad de ambas instituciones por ayudarme con sus comentarios y observaciones críticas.

Finalmente, varios de mis amigos y familiares han leído todo el manuscrito o partes de él, siempre con resultados positivos para la obra. Tengo que decir que ninguna de estas personas es responsable de los errores que puedan quedar: profesores Ned Block, Susan Carey Block, George Boolos, Noam Chomsky, Janet Dean Fodor, Jerrold Katz, Edward Martin y George Miller. Estoy especialmente agradecido a Mr. Georges Rey, que leyó el manuscrito con gran detenimiento y me ayudó con sus críticas y consejos; y a Mrs. Cornelia Parkes, que me ayudó en la bibliografía.

La segunda mitad de la Introducción al libro es una versión, ligeramente revisada, de un artículo titulado «Special Sciences», que apareció por primera vez en *Synthese* (Fodor, 1974). Agradezco la autorización para volver a publicarlo. El material citado de los capítulos uno y dos y la conclusión de *La construcción de la realidad en el niño,* de Jean Piaget (traducido al inglés por Marjorie Cooke), es propiedad de Basic Publishers, New York (1954), y lo utilizamos con su autorización. Otras citas están hechas con autorización de D. Riedel Publishing Co.; Penguin Books Ltd., John Wiley and Sons Inc., The Humanities Press Inc., y Routledge and Kagan Paul Ltd.

PREFACIO A LA EDICION CASTELLANA

La primera edición (en inglés) de *The Language of Thought* está fechada en 1975. Habían pasado casi veinte años desde la publicación (1957) de la famosa monografía de Noam Chomsky, *Syntactic Structures,* y quice años desde la aparición (1960) de *Plans and the Structure of Behavior,* de George Miller, Eugene Galanter y Karl Pribram. *El lenguaje del pensamiento* (LDP) aceptaba en gran parte la idea que se daba de la mente cognitiva en estos dos libros. De la misma manera, LDP es bastante posterior a sus principales adversarios. *The Concept of Mind,* de Ryle —cuyo conductismo es ampliamente criticado en el capítulo introductorio— se publicó en 1949; el trabajo de Paul Oppenheim y Hilary Putnam, «The Unity of Science as a Working Hypothesis», que constituye el otro gran blanco de la Introducción, apareció en 1958.

Estas reflexiones cronológicas refuerzan una idea expresamente reflejada en el primer prefacio a LDP; el objetivo del libro no era convertirse en una obra pionera, sino hacer un intento de «consolidación». Por una parte, en 1975 el funcionalismo iba ya camino de convertirse en la doctrina «oficial» de la filosofía anglo-americana de la mente, sustituyendo a las diversas formas de reduccionismo contra las que van dirigidos fundamentalmente los pasajes metodológicos de LDP. Por la otra, LDP llegaba tras más de una década de trabajos en el campo de la lingüística generativa y de la psicología cognitiva. Quizá el resultado más importante de tales trabajos había sido familiarizar a la comunidad investigadora con dos de las ideas fundamentales de lo que ahora se llama «Ciencia cognitiva»: que los estados mentales son típicamente *representacionales* y que los procesos mentales son típicamente *computacionales.* LDP fue ante todo un intento de integrar y elaborar estas dos ideas; de aceptarlas literalmente y con toda seriedad como base para una teoría de la mente cognitiva.

Decir que los estados mentales son representacionales es decir que tienen contenido (así, por ejemplo, la creencia de que mañana va a llover tiene el contenido *mañana lloverá*). Decir que los procesos mentales son computacionales es decir que son —en algún sentido de estos términos tan castigados— «formales» o «sintácticos». El

enigma que LDP intentaba resolver era éste: «*¿cómo es posible* que los estados mentales tengan contenido y que los procesos mentales sean sintácticos?».

La solución propuesta por LDP estaba inspirada no sólo en los supuestos implícitos en la naciente ciencia cognitiva, sino también en la epistemología clásica de los siglos XVII y XVIII. Podríamos resumirla en estas palabras: Los procesos mentales son secuencias causales de estados mentales. Los estados mentales son relaciones entre organismos y *símbolos* mentales (en LDP se denomina a los símbolos mentales con el término técnico de «representaciones mentales»; Locke y Hume los habrían llamado «ideas»). Dada la postulación de los símbolos mentales, encajan muchas cosas. Por definición, los símbolos tienen contenido semántico y forma sintáctica; eso es precisamente un símbolo. Por eso, los estados mentales son representacionales porque son relaciones con símbolos mentales y los símbolos mentales tienen contenido semántico. Y los procesos mentales son sintácticos porque operan sobre los símbolos mentales y los símbolos mentales tienen forma sintáctica. La mente cognitiva aparece como una máquina para manipular representaciones; una especie de ordenador, pero hecho con proteínas en vez de con silicona.

LDP creía que había que aceptar esta solución, aunque sólo fuera porque no había otra alternativa. Más que demostrar la teoría de la representación mental lo que trataba de hacer era precisarla, explicitando sus consecuencias. Algunas de las doctrinas predilectas de LDP han adquirido cierta notoriedad —en especial el nativismo comprensivo adoptado en el capítulo 2—, pero, de hecho, no son más que la consecuencia lógica de una aceptación escrupulosa de la explicación computacional/representacional de la mente cognitiva. Si, por ejemplo, el aprendizaje es un proceso computacional —un proceso definido sobre símbolos mentales—, estos símbolos mentales tienen que proceder de alguna parte; si son aprendidos, los símbolos que median *ese* aprendizaje deben proceder, a su vez, de algún lugar. Tarde o temprano, este retroceso deberá detenerse. Cualquiera que sea el punto en que nos detengamos, se acepta ipso facto que es *no* aprendido; y por tanto que es innato. La idea de LDP era que, dado que el nativismo comprensivo forma parte del coste que hay que pagar para tener una teoría computacional/representacional de la mente, no queda otro remedio que pagar tal precio; efectivamente, hacia 1975, la teoría computacional/representacional de la mente no tenía competidores serios. En mi opinión, sigue sin tenerlos.

¿Qué opinión merece todo esto, visto retrospectivamente? En cierto sentido, el estudio de la mente ha avanzado en gran parte en la línea propuesta por LDP. Una premisa metodológica de LDP era que la postulación de representaciones mentales era la única manera de hacer científicamente respetable a la psicología de las «actitudes proposicionales», inspirada en el sentido común: la psicología o era computacional o no era psicología. Esta forma de entender las alternativas existentes sigue siendo válida diez años más tarde. Los críticos filosóficos contemporáneos, que dudan sobre la teoría computacional/representacional, suelen hablar de lo mental dentro de una línea «eliminativista». Es decir, dudan de la posibilidad de una ciencia que suponga que la conducta cognitiva más elevada es resultado de los deseos y creencias de un organismo; y prevén la futura sustitución de las explicaciones psicológicas de la conducta por teorías formuladas con el lenguaje de la neurología o de la biología. Pero reina la misma confusión que hace diez años sobre la realización concreta de es-

te programa eliminativista y sobre la concepción filosófica de la acción humana que en él se implica.

También los desarrollos empíricos parecen haber seguido en gran parte el camino previsto. LDP suponía que el progreso de la lingüística y de la psicología cognitiva daría lugar a teorías, cada vez más desarrolladas y confirmadas, sobre la naturaleza de las representaciones mentales y sobre el carácter de los procesos computacionales que las incluyen. Así ha sido, con resultados espectaculares no sólo en el estudio del lenguaje, sino también en la psicología de las imágenes mentales y de la psicofísica visual. En general, el cuadro del programa de investigación cognitiva ofrecido por LDP sigue siendo, en mi opinión, exacto (aunque los detalles psicológicos y lingüísticos que considera LDP serían distintos si escribiéramos ahora los capítulos 3 y 4). Me produce especial satisfacción la insistencia de LDP en una posible *pluralidad* de códigos mentales (véase sobre todo el capítulo 4). Prefigura un enfoque «modular» de las funciones mentales, idea que la ciencia cognitiva de nuestros días parece estar tomándose muy en serio.

Me gustaría poder decir que los progresos de la ciencia cognitiva expuestos en LDP contienen al menos las líneas maestras de una explicación general y satisfactoria sobre la mente cognitiva; que se entienden, al menos en sus grandes líneas, los principales problemas teóricos y filosóficos y que lo único que queda por hacer es completar los detalles empíricos. Debo poner en guardia al lector frente a este optimismo fácil. Creo que LDP es una buena prueba de la necesidad de postular representaciones mentales en toda explicación de los procesos mentales que podamos concebir. Pero LDP guarda un silencio total —por no decir ensordecedor— sobre un problema fundamental con que la psicología computacional tendrá que enfrentarse en algún momento: ¿De dónde proceden las propiedades semánticas de los símbolos mentales? ¿Cómo consiguen representar las representaciones mentales? Los supuestos metafísicos de LDP son materialistas; se da por supuesto que las representaciones mentales son objetos físicos. Pero en ese caso, ¿cómo pueden tener propiedades semánticas los objetos físicos? ¿Cómo pueden versar los unos sobre los otros? Este problema procede en línea directa del famoso «problema de la intencionalidad» de Brentano; en realidad, es la forma que adopta el problema de Brentano en el contexto de una teoría representacional de la mente. En la actual situación, no tenemos la menor idea de cómo solucionarlo, aunque debemos intentarlo si queremos tener una ciencia cognitiva que sea al mismo tiempo materialista y computacional.

Visto retrospectivamente, quizá el principal logro de LDP haya sido demostrar lo importante que es progresar en este problema. Siempre ha estado claro que no puede haber teoría del *lenguaje* sin una serie de símbolos. Lo que LDP insinúa —quizá de forma sorprendente— es que tampoco puede haber una teoría de la *mente* sin una teoría de los símbolos.

JERRY A. FODOR
Cambridge, Mass., 1984

Introducción
DOS CLASES DE REDUCCIONISMO

El hombre que ríe es aquel
que todavía no se ha enterado de
la terrible noticia.

BERTOLT BRECHT

En este libro me propongo tratar de algunos de los aspectos de la teoría de los procesos mentales. Muchos lectores pueden considerar que la elección del tema es poco afortunada: bien porque piensan que no existen tales procesos y por lo tanto no es posible hablar de ellos, o bien porque creen que no hay ninguna teoría sobre ellos cuyos aspectos puedan ser objeto de discusión. La segunda de estas consideraciones es importante, y nos ocuparemos de ella en el conjunto del texto. Después de todo, la mejor demostración de que es posible hacer psicología especulativa es hacerla. Pero soy consciente de que la desconfianza con que muchos filósofos, y muchos psicólogos con tendencias filosóficas, consideran el tipo de estudio que voy a emprender, procede de algo que no es simplemente la estima y envidia que les producen las publicaciones de carácter empírico. En este capítulo nos ocuparemos de las fuentes de esta sospecha.

La integridad de la teoría psicológica se ha visto amenazada desde siempre por dos clases de reduccionismo, cada uno de los cuales viciaría en raíz la pretensión del psicólogo de estudiar los fenómenos mentales. Para los que se dejan influir por la tradición del conductismo lógico, estos fenómenos no merecen ninguna consideración ontológica distinta de los hechos de conducta que explican las teorías psicológicas. De esta forma la psicología se ve privada de sus términos teóricos a no ser que los interpretemos como locuciones inventadas y sin contenido de las que llegarán a ofrecerse reducciones conductuales. A todos los efectos, esto quiere decir que los psicólogos pueden únicamente ofrecer explicaciones metodológicamente acreditadas de los aspectos de la conducta que son resultado de variables ambientales.

No es de extrañar que muchos psicólogos hayan encontrado demasiado restrictiva e intolerable esta metodología: la aportación de los estados internos del organismo a la producción de su propia conducta parece estar suficientemente por encima de toda disputa, teniendo en cuenta la espontaneidad y libertad ante el posible control ambiental que se aprecia muchas veces en la conducta. Por eso, el conductismo nos invi-

ta a negar lo que está fuera de toda disputa, pero, en realidad, no hace falta que lo hagamos; existe una segunda posibilidad, que es la que se suele adoptar. Podemos reconocer que la conducta es en gran parte consecuencia de procesos orgánicos en la medida en que tengamos en cuenta que estos procesos *son* orgánicos: es decir, que son procesos fisiológicos localizados, probablemente, en los sistemas nerviosos de los organismos. De esta manera la psicología puede evitar la reducción conductual tomando opción por la reducción fisiológica, pero debe optar en uno u otro sentido.

En uno y otro caso el psicólogo sale perdiendo. En tanto en cuanto las explicaciones psicológicas admiten un vocabulario teórico, se trata del vocabulario de una ciencia *diferente* (neurología o fisiología). En tanto en cuanto *existen* leyes sobre las formas en que la conducta depende de los procesos internos, es el neurólogo o el fisiólogo quien llegará, con el tiempo, a formularlas. Cualquiera que sea la elección de los psicólogos entre las reducciones disponibles, su disciplina se queda sin un objeto material que la pertenezca en exclusiva. Lo más a que puede aspirar un psicólogo en activo es a una existencia provisional y difícil entre los cuernos de este dilema y a ser (solamente) tolerado por los colegas de las ciencias «duras».

Sin embargo, yo creo que se trata de un falso dilema. No veo ninguna razón convincente que impida a la ciencia tratar de mostrar la dependencia de la conducta de un organismo en relación a sus estados internos, y tampoco sé de ninguna razón por la que aquella ciencia que consiga hacerlo tenga que ser reducible a la ciencia del cerebro; no, al menos, en el sentido de una reducción que implicara que las teorías psicológicas pueden de algunas maneras ser *reemplazadas* por sus equivalentes fisiológicas. Voy a intentar, en este capítulo introductorio, hacer ver que los cuernos del dilema están en realidad embotados. Con ello espero quitar la base a varios de los argumentos que se suelen presentar normalmente contra algunos tipos de explicaciones psicológicas que, en capítulos posteriores, yo voy a considerar con toda seriedad.

EL CONDUCTISMO LOGICO

Muchos filósofos, y algunos científicos, parecen mantener que las clases de teorías sostenidas ahora en general por los psicólogos cognitivos no tienen posibilidades de iluminar el carácter de los procesos mentales. Estas teorías, se afirma, adoptan una actitud sobre la explicación psicológica que es, tal como se ha demostrado, fundamentalmente incoherente. Dicho sin rodeos, Ryle y Wittgenstein habrían matado esta clase de psicología en un momento que estaría en torno al año 1945, y no tiene sentido hacer especulaciones sobre las posibilidades del difunto.

No voy a tratar de hacer una refutación detallada de esta opinión. Si es cierto que la tradición wittgensteiniana representa un ataque coherente contra la metodología de la psicología cognitiva actual, sería un ataque que depende de un complejo de presuposiciones sobre la naturaleza de la explicación, la categoría ontológica de las entidades teóricas, y las condiciones a priori de la posibilidad de comunicación lingüística. Para resistir de frente este ataque sería preciso demostrar —de hecho, yo creo que es cierto— que estas suposiciones, en la medida en que son claras, no están justificadas. Pero esto es una tarea que necesitaría todo un libro, y un libro que nunca he tenido la tentación de escribir. Lo mejor que puedo hacer aquí es esbozar una defensa preli-

minar de los compromisos metodológicos implícitos en la forma de elaboración teórica de contenido psicológico de que me voy a ocupar fundamentalmente. En tanto en cuanto que estos compromisos difieren de lo que muchos filósofos han estado dispuestos a aceptar, este simple esquema de su defensa puede resultar revelador.

Entre los muchos pasajes de *The Concept of Mind,* de Ryle (1949), que merecen ser considerados con gran atención, hay uno (p. 33) en que se ponen las cartas sobre la mesa, desde luego en mayor medida de lo que es habitual. Ryle está tratando la cuestión: «¿Qué es lo que hace que las payasadas de un payaso sean inteligentes (ingeniosas, agudas, etc.)?». La doctrina que trata de rechazar se podría formular así: Lo que hace que sean inteligentes es el hecho de que son consecuencia de ciertas operaciones mentales (computaciones, cálculos) propias del payaso y responsables en sentido causal de la producción de la conducta del mismo. Si estas operaciones hubieran sido diferentes, en ese caso (sigue afirmando tal doctrina) o bien las payasadas habrían carecido de ingenio o al menos habrían sido payasadas inteligentes pero de una naturaleza distinta. En resumidas cuentas, las payasadas del payaso eran inteligentes en la forma en que lo eran debido a que las operaciones mentales de que dependían causalmente tenían un carácter determinado, y no otro. Y, aunque Ryle no lo dice, en esta doctrina se presupone implícitamente que un psicólogo interesado en explicar el éxito de la actuación del payaso se vería, ipso facto, en la obligación de decir cuáles eran esas operaciones y cómo se relacionaban precisamente con las volteretas que veía claramente el público.

Estrictamente hablando, esto no es una teoría única, sino un conjunto de teorías estrechamente vinculadas. En concreto, se pueden distinguir al menos tres afirmaciones sobre el carácter de los hechos de que parece depender causalmente la conducta del payaso:

1. Que algunos de ellos son hechos mentales;
2. Que algunos (o todos) hechos mentales son propios del payaso al menos en el sentido de que normalmente no son observados por alguien que observe la actuación del payaso y, quizá, también en el sentido más fuerte de que, por principio, no son observables para nadie, excepto para el payaso;
3. Que es el hecho de que la conducta estuviera causada por tales hechos lo que la convierte en el tipo de conducta que es de hecho; que la conducta inteligente *es* inteligente porque tiene esa etiología concreta.

Quiero distinguir entre estas doctrinas porque un psicólogo podría aceptar las clases de teoría que no gustan a Ryle sin por ello tener que aceptar todas las implicaciones de lo que Ryle llama «cartesianismo». Por ejemplo, Ryle presupone (cosa que no harían la mayoría de los psicólogos que adoptan una opinión realista de las realidades designadas por los términos mentales en las teorías psicológicas) que para ser mentalista hay que ser dualista; en concreto, que el mentalismo y el materialismo se excluyen mutuamente. En otra obra he defendido que el pecado original de la tradición wittgensteiniana está en confundir el mentalismo con el dualismo (cf. Fodor, 1968, especialmente el capítulo 2). Aquí nos limitaremos a señalar que uno de los resultados de esta confusión es la tendencia a considerar que las opciones del dualismo y el conductismo agotan las posibilidades de la filosofía de la mente.

De la misma manera, me parece, podrían aceptarse puntos de vista como el punto

3, sin tener que aceptar una interpretación doctrinaria del punto 2. Puede que algunos de los procesos mentales que son causalmente responsables de la conducta del payaso sean, de facto, inobservables para el público. En este sentido, también podría ser que algunos de estos procesos sean, de facto, inobservables para el payaso. Pero no parece que haya nada, en el proyecto de explicar la conducta por referencia a procesos mentales, que exija mantenerse fiel al carácter privado de la epistemología en el sentido tradicional de esa noción. En realidad, para bien o para mal, un materialista *no puede* aceptar este compromiso pues su opinión es que los hechos mentales son especies de hechos físicos, y los hechos físicos son públicamente observables, al menos en principio[1,2].

Es claro que, incluso teniendo en cuenta estas precauciones, Ryle no cree en la posibilidad de que resulte verdadera esta forma de explicación. En ella se dice que lo que hace que las payasadas del payaso sean inteligentes es el hecho de que son resultado de un tipo determinado de causa. Pero lo que, según el punto de vista de Ryle, *hace* verdaderamente que sean inteligentes es algo completamente distinto: Por ejemplo, el hecho de que ocurran externamente, donde el público pueda verlas; el hecho de que las cosas que hace el payaso no son las cosas que el público esperaba que hiciera; el hecho de que el hombre a quien tiró la tarta fuera vestido de etiqueta, etc..

Hay que tener en cuenta dos puntos. En primer lugar, ninguno de *estos* hechos es algo privado del payaso, en ningún sentido. Ni siquiera son privados en el sentido de que sean hechos sobre cosas que ocurren en el sistema nervioso del payaso. Por el contrario, lo que hace que las payasadas del payaso sean inteligentes son precisamente los aspectos *públicos* de su actuación; precisamente las cosas que la audiencia *puede* ver. La segunda cosa a tener en cuenta es que lo que hace que las payasadas sean inteligentes no es el carácter de las *causas* de la conducta del payaso, sino más bien el carácter de la conducta misma. Influye en que la voltereta fuera inteligente el hecho de que ocurriera cuando nadie la esperaba, pero el que ocurriera cuando no se la esperaba no era indudablemente una de sus causas en ninguna interpretación posible de

---

[1] Los puristas observarán que este último punto depende de la (razonable) suposición de que el contexto «es públicamente observable al menos en principio» transparente a la substitutividad de los idénticos.

[2] Podría replicarse que si admitimos la posibilidad de que los hechos mentales sean hechos físicos, de que algunos hechos mentales sean inconscientes, y de que ningún hecho mental sea esencialmente privado, habremos debilitado el término «mental» hasta el punto de quitarle toda su fuerza. Naturalmente, es cierto que la noción misma de hecho mental se especifica frecuentemente de maneras que presuponen el dualismo y/o una doctrina fuerte de la intimidad epistemológica. Lo que no está claro, sin embargo, es *para* qué queremos, en primer lugar, una definición de «hecho mental».

Desde luego que no, en cualquier caso, para que sea posible hacer psicología de forma respetable. *Pre*teóricamente identificamos los hechos mentales por referencia a casos claros. *Post*-teóricamente basta con identificarlos en cuanto que caen dentro de las leyes psicológicas. Esta caracterización constituye, lógicamente, una petición de principio pues se basa en una distinción entre leyes psicológicas y todas las demás que no se explica. No obstante, lo que se trata de recordar ahora es que nosotros no nos encontramos en mejor situación frente a nociones como las de fenómeno químico (o meteorológico, o geológico..., etc.), situación que no impide la búsqueda racional de la química. Un fenómeno químico es aquel que cae dentro de las leyes químicas; leyes químicas son las que se derivan de las teorías químicas (supuestamente desarrolladas); teorías químicas son las teorías que se dan en química; y la química, como todas las demás ciencias especiales, se individualiza en gran parte a posteriori y por referencia a sus problemas y predicados característicos. (Por ejemplo, la química es la ciencia que se ocupa de las propiedades combinatorias de los elementos, el análisis y síntesis de los compuestos, etc.). ¿Por qué, exactamente, no basta con esto?

«causa». Resumiendo, lo que hace que las payasadas sean inteligentes no es un hecho distinto de, y causalmente responsable de, la conducta que produce el payaso. A fortiori, no es un hecho mental anterior al revolcón. Indudablemente, si el programa mentalista implica la identificación y caracterización de este hecho, dicho programa está condenado al fracaso desde el primer momento.

Lo sentimos por la psicología de las payasadas inteligentes. Habíamos supuesto que los psicólogos identificarían las causas (mentales) de que dependían las payasadas inteligentes y *por lo tanto* responderían a la pregunta: «¿Qué es lo que hace que las payasadas sean inteligentes?». Lo único que parece quedar del intento son las aliteraciones. Pero Ryle no reduce su utilización de esta forma de argumentar al intento de echar por tierra la psicología de los payasos. Se realizan movimientos muy semejantes para demostrar que la psicología de la percepción es un embrollo, pues lo que hace que algo sea, por ejemplo, reconocimiento de un petirrojo o de una melodía no es la existencia de uno u otro hecho mental, sino más bien el hecho de que lo que se dijo que era un petirrojo fuera de hecho un petirrojo, y que lo que se consideró como una versión del «Lillibulero» lo fuera realmente. Resulta francamente difícil pensar en un área de la psicología cognitiva en que no se pueda aplicar esta forma de argumentación o donde no la aplique Ryle. De hecho, quizá la afirmación *central* de Ryle es que las teorías psicológicas «cartesianas» (es decir, mentalistas) tratan lo que es en realidad una relación *lógica* entre aspectos de un único fenómeno como si se tratara en verdad de una relación causal entre parejas de hechos distintos. Es esta tendencia a ofrecer respuestas mecanicistas a cuestiones conceptuales lo que, según Ryle, lleva a los mentalistas a orgías de hipóstasis lamentables: es decir, a intentar explicar la conducta por referencia a los mecanismos psicológicos subyacentes[3].

Si resulta que esto *es* un error, mi situación es comprometida. La suposición que va a servir de base a toda mi exposición es que tales explicaciones, por muy elevado que sea el número de veces en que resultan ser erróneas empíricamente, son, en principio, metodológicamente impecables. Lo que me propongo hacer a lo largo del libro es tomar estas explicaciones completamente en serio e intentar trazar al menos un esbozo de la imagen general de la vida mental a que nos llevan. Por eso, algo habrá que hacer para responder al argumento de Ryle. Por de pronto, vamos a cambiar el ejemplo.

Pensemos en esta pregunta: «¿Qué hace que Wheaties sea el desayuno de los campeones?». («Wheaties», en caso de que alguien no haya oído la palabra, es, o son, una clase de cereales envasados. Los detalles no tienen importancia). Como es fácil comprobar, se podrían presentar al menos dos tipos de respuestas[4]. Una de es-

---

[3] Ryle no utiliza el término «criterio»: Sin embargo, la línea argumental que acabamos de mencionar relaciona estrechamente la obra de Ryle con la tradición criteriológica de la filosofía post-wittgensteiniana de la mente. Aproximadamente, lo que, desde el punto de vista de Ryle, «hace» que *a* sea *F* es que *a* posea las propiedades que tienen carácter de criterio para la aplicación de «*F*» a *x*.

[4] Cuando leo «¿Qué hace que Wheaties sea el desayuno de los campeones?» yo entiendo que se pregunta «¿Qué hace que (algunos, muchos, tantos de) los que comen Wheaties sean campeones?» y no «¿Qué hace que (algunos, muchos, tantos de) los campeones coman Wheaties?». Esta última pregunta sugiere las razones que dan los campeones para comer Wheaties; y aunque es *posible* que hagan referencia a las propiedades que tienen los Wheaties en virtud de las cuales los que los comen llegan a ser campeones, no es

tas respuestas, que pertenece a lo que voy a denominar como «historia causal», iría más o menos en esta línea: «Lo que hace de Wheaties el desayuno de los campeones son las vitaminas y minerales vigorizantes que contienen»; o «Son los hidratos de carbono de los Wheaties, que dan la energía necesaria para los esfuerzos requeridos por el salto de altura»; o «Es la elasticidad de todas las pequeñas moléculas de Wheaties lo que da a los que toman Wheaties su coeficiente o capacidad de recuperación excepcionalmente altos», etc.

No tiene importancia para mi argumentación si estos modelos de respuesta son verdaderos o no. Lo que sí *es* esencial es que una u otra de las historias causales debe ser cierta para que los Wheaties *sean* realmente el desayuno de los campeones, tal como se afirma en dicha fórmula. Las respuestas proponen historias causales en la medida en que tratan de especificar las propiedades de Wheaties que pueden estar causalmente relacionadas con los procesos que hacen campeones a los que comen Wheaties. En términos aproximados, estas respuestas sugieren valores provisionales de *P* en el siguiente esquema de explicación: «*P* es causa ((*x* come Wheaties) da lugar a (*x* se convierte en un campeón)) para un número significativamente elevado de valores de *x*». Supongo que, si los Wheaties hacen realmente campeones a aquellos que los comen, tiene que haber al menos un valor de *P* que haga que este esquema sea verdadero. Como dicha suposición es sencillamente una negativa a admitir la teoría de que los Wheaties son milagrosos, no hay razón para discutirla.

He señalado antes que hay otra forma de respuesta que se puede aplicar adecuadamente a «¿Qué hace que Wheaties sea el desayuno de los campeones?». Las respuestas de esta segunda categoría pertenecerían a lo que llamaremos «historia conceptual». En el caso que nos ocupa, podemos contar la historia conceptual con cierta precisión: Lo que hace de Wheaties el desayuno de los campeones es el hecho de que es tomado (para desayunar) por un número no despreciable de campeones. Supongo que esta es una condición conceptualmente necesaria y suficiente para que *algo* sea el desayuno de los campeones[5]; en cuanto tal, agota prácticamente la historia conceptual de Wheaties.

Lo que hay que tener en cuenta de todo esto es que las respuestas que pertenecen a la historia conceptual no pertenecen por norma a la historia causal, y viceversa[6].

---

necesario que así sea. Así, una respuesta plausible a la segunda pregunta que *no* es respuesta plausible para la primera sería: «saben bien».

No estoy seguro de cuál es la pregunta que tienen presente los hombres de Wheaties cuando preguntan «¿Qué hace de Wheaties el desayuno de los campeones?» en tono retórico, como, creo yo, suelen hacer. Gran parte de su propaganda consiste en dar publicidad a las afirmaciones de los campeones en el sentido de que ellos (los campeones) consumen Wheaties. Si, como puede ocurrir, estas afirmaciones se ofrecen como argumentos en favor de la presuposición de la pregunta en su *primera* interpretación (es decir, que los Wheaties *tienen* algo que hace campeones a los que los comen), resultaría que la General Mills o ha utilizado indebidamente el método de las diferencias o cometido la falacia de la afirmación del consecuente.

Se puede hacer filosofía de cualquier cosa.

[5] Esto no es totalmente cierto. El ser tomado para desayunar por un número no despreciable de campeones es condición conceptualmente necesaria y suficiente para que algo sea *un* desayuno de campeones (cf. Russell, 1905). En adelante, resistiré a esta forma de pedantería siempre que me sea posible.

[6] Las excepciones son interesantes. Implican casos en que las condiciones conceptuales para que algo sea una cosa de una determinada clase incluyen la necesidad de que tenga, o sea, una cierta clase de causa. Supongo, por ejemplo, que es una verdad conceptual el que no se puede considerar que algo es una pelea

En concreto, el que sean ingeridos por un número no despreciable de campeones no es *causa* de que Wheaties sea el desayuno de los campeones; como tampoco el hecho de que se produzca inesperadamente es causa de que la voltereta del payaso sea graciosa. Más bien, lo que tenemos en ambos casos son ejemplos de conexiones conceptuales (más o menos rigurosas). El ser comido por un número no despreciable de campeones y el ser inesperado pertenecen, respectivamente, a los análisis de «ser el desayuno de los campeones» y de «ser gracioso», con la excepción de que, en el primer caso, tenemos algo que se aproxima a una condición lógicamente necesaria y suficiente y, en el segundo, es evidente que no se da esa circunstancia[7].

La idea de conexión conceptual es un caso patente de miasma filosófico; tanto más si se afirma (como suelen hacer Wittgenstein y sus seguidores) que hay tipos de conexiones conceptuales que, al menos en principio, no se pueden explicar desde el punto de vista de las nociones de condiciones lógicamente necesarias y/o suficientes. Sin embargo, lo que queremos subrayar en este momento es que en *cualquier* interpretación razonable de la conexión conceptual, los Wheaties demuestran que *tanto* la historia causal *como* la conceptual pueden ser respuestas simultáneamente verdaderas y distintas a preguntas del tipo: «¿Qué hace que (un) *x* sea (un) *F*?». Para decirlo en pocas palabras, el dietético que aparece en la televisión para explicar que los Wheaties son el desayuno de los campeones porque contienen vitaminas, no se ve refutado por el filósofo que observa (aunque no por televisión, generalmente) que Wheaties es el desayuno de los campeones porque los campeones lo toman para desayunar. El especialista en dietética, al decir lo que dice, no supone que sus observaciones expresen, o puedan sustituir a, las verdades conceptuales pertinentes. El filósofo, al decir lo que dice, no debería suponer que sus observaciones expresen, o puedan sustituir a, las explicaciones causales pertinentes.

Supongamos, en general, que *C* es una condición conceptualmente suficiente para tener la propiedad *P*, y supongamos que un determinado individuo *a* cumple, de hecho, con *C*, de manera que «*Pa*» sea una afirmación contingente verdadera en relación a *a*. Entonces: (a) resulta normalmente pertinente exigir una explicación causal/mecanicista del hecho de que «*Pa*» *es* verdad; (b) esta explicación constituirá normalmente una (aspirante a) respuesta a la pregunta: «¿Qué es lo que hace que *a* tenga la propiedad *P*?»; (c) la referencia al hecho de que *a* cumple con *C no* constituirá normalmente una explicación causal/mecanicista del hecho de que *a* tenga la

---

entre borrachos a no ser que la borrachera de los alborotadores haya contribuido causalmente a producir la pelea. Otros ejemplos: los virus de la gripe, las lágrimas de rabia, los suicidios, los tartamudeos nerviosos, etc. En realidad, podemos imaginarnos un análisis de «el desayuno de los campeones» que lo convirtiera en uno de estos casos; es decir, un análisis que diga que es lógicamente necesario que el desayuno de los campeones sea (no sólo lo que los campeones toman para desayunar sino también) lo que los campeones toman para desayunar y que es causalmente responsable de que sean campeones. Pero ¡ya vale!

[7] A propósito, no es ningún accidente que este último caso sea incompleto. La situación habitual es que las condiciones lógicamente necesarias y suficientes para la atribución de un estado mental a un organismo se refieren no solamente a las variables ambientales, sino a otros estados mentales de dicho organismo. (Por ejemplo, *saber* que *P* es *creer* que *P* y cumplir algunas condiciones más; *ser codicioso* es estar dispuesto a *experimentar satisfacción* al conseguir, o ante la perspectiva de conseguir, más de lo que corresponde, etc.). La fe en que *debe* haber una forma de salir de esta red de términos mentales interdependientes —que se puede llegar a atribuciones conductuales puras con sólo progresar suficientemente en el análisis— no se ve apoyada, por lo que yo sé, ni por argumentos ni por ejemplos.

propiedad *P;* aunque, (d) las referencias al hecho de que *a* cumple con *C* puede constituir un cierto modelo (diferente) de respuesta a «¿Qué hace que *"Pa"* sea verdad?». Supongo que, excluyendo la imprecisión de la noción de una conexión conceptual (y, a ese respecto, la ambigüedad de la noción de explicación causal), este modelo se refiere al caso especial en que *C* consiste en la propiedad de ser inesperado, *a* es un revolcón, y «*Pa*» es la afirmación de que *a* estuvo gracioso.

Para decirlo en la forma más general que me es posible, aun cuando los conductistas tuvieran razón al suponer que las condiciones lógicamente necesarias y suficientes para que la conducta sea de una determinada categoría se puedan presentar (solamente) en términos de estímulo y respuesta, tal hecho no desmentiría en lo más mínimo la afirmación del mentalista de que la *causalidad* de la conducta está determinada por, y se puede explicar desde el punto de vista de, los estados internos del organismo. Por lo que yo sé, la escuela filosófica del conductismo «lógico» no ofrece nada que se parezca a un argumento que nos obligue a creer que tal afirmación es falsa. Y el hecho de que la psicología conductista no llegue a presentar ni siquiera una primera aproximación a una teoría plausible del conocimiento nos hace pensar que la afirmación del mentalista puede ser perfectamente verdadera.

Los argumentos que hemos estado considerando van dirigidos contra una clase de reduccionismo que trata de demostrar, de una u otra manera, que los hechos mentales a que recurren las explicaciones psicológicas no pueden ser antecedentes causales de los hechos de conducta que tratan de explicar las teorías psicológicas; a fortiori, que las afirmaciones que atribuyen la inteligencia de una actuación a la calidad de las funciones cerebrales del agente no pueden ser etiológicas. El tema recurrente en esta forma de reduccionismo es la alegación de una conexión conceptual entre los predicados conductuales y mentales en los ejemplos característicos de las explicaciones psicológicas. Es a partir de la existencia de esta conexión de donde se deduce la categoría de segunda clase de los hechos mentales.

A estas alturas debería ya haber quedado claro que no creo que se pueda mantener este tipo de argumento. Por eso supondré, de aquí en adelante, que los psicólogos se dedican a ofrecer teorías sobre los hechos que intervienen causalmente en la producción de la conducta y que los psicólogos cognitivos se dedican a ofrecer teorías sobre los hechos que intervienen causalmente en la producción de conducta inteligente. No existe, por supuesto, ninguna garantía de que se pueda realizar este juego. Es perfectamente posible que las clases de conceptos en que se elaboran las teorías psicológicas actuales resulten, a la larga, poco adecuadas para la explicación de la conducta. Por eso, es perfectamente concebible que los procesos que intervienen en la producción de la conducta sean demasiado complicados para que nadie llegue a entenderlos. Nadie puede demostrar, a priori, que un programa de investigación empírica va a resultar provechoso. Lo que yo quiero dejar claro es únicamente que los conductistas lógicos no han presentado ninguna razón a priori que haga suponer que el programa mentalista no lo vaya a ser.

Sin embargo, si no queremos que los hechos mentales queden reducidos a hechos conductuales, ¿qué es lo que debemos decir sobre su categoría ontológica? Creo que es muy probable que todas las causas orgánicas de la conducta sean fisiológicas, y por lo tanto que los hechos mentales pueden llegar a tener descripciones verdaderas en el vocabulario de una fisiología supuestamente terminada. Pero creo que no tiene

ningún interés el que yo lo crea. En concreto, no supongo siquiera que comience a derivarse de esta clase de materialismo que toda rama de la fisiología proporcione o pueda proporcionar el vocabulario adecuado para la construcción de teorías psicológicas. La probabilidad de que los hechos psicológicos sean hechos fisiológicos no implica la reducibilidad de la psicología a fisiología, a pesar de que haya muchos filósofos y fisiólogos que opinen lo contrario. Para ver por qué son así las cosas hace falta una exposición bastante amplia de toda la noción de reducción intercientífica, noción que ha contribuido tanto como la que más —si exceptuamos, quizá, el criterio de verificabilidad del significado— a oscurecer la metodología de la psicología.

EL REDUCCIONISMO FISIOLOGICO

Una tesis característica de la filosofía positivista de la ciencia es que todas las teorías verdaderas de las ciencias especiales deberían reducirse, «a la larga», a teorías físicas. Esta tesis pretende ser empírica, y parte de las pruebas en que se basa las proporcionan éxitos científicos tales como la teoría molecular del calor y la explicación física del enlace químico. Pero la popularidad filosófica del programa reduccionista no se puede explicar por referencia a estos logros únicamente. El desarrollo de la ciencia ha sido testigo de la proliferación de disciplinas especializadas al menos con la misma frecuencia que ha sido testigo de su eliminación, por lo que la difusión del entusiasmo a favor de la opinión de que eon el tiempo sólo habrá física no puede ser una mera inducción de los anteriores éxitos reduccionistas.

Creo que muchos filósofos que aceptan el reduccionismo lo hacen porque quieren respaldar la generalidad de la física frente a las ciencias especiales: en términos aproximados, el punto de vista de que todos los hechos que vendrán a caer dentro de las leyes de cualquier ciencia son hechos físicos y por lo tanto quedarán bajo las leyes de la física[8]. Para estos filósofos, da la impresión de que decir que la física es una ciencia básica y decir que las teorías de las ciencias especiales deben reducirse a teorías físicas es decir lo mismo de dos maneras diferentes, por lo que la última doctrina se ha convertido en la interpretación consagrada de la primera.

En las páginas siguientes intentaré demostrar que esto constituye una confusión considerable. Lo que se ha venido denominando tradicionalmente «la unidad de la ciencia» es una tesis mucho más fuerte y mucho menos plausible que la generalidad de la física. Si esto es cierto, tiene su importancia. Aunque el reduccionismo es una doctrina empírica, está concebido para desempeñar un papel regulador en la práctica científica. Se considera que la reductibilidad a la física es una *limitación* de la aceptabilidad de las teorías en las ciencias especiales, con la consecuencia curiosa de que cuanto mayor éxito tienen las ciencias especiales más llamadas están a desaparecer. Los problemas metodológicos sobre la psicología, en concreto, se presentan de esta manera: Se considera que la suposición de que el objeto material de la psicología forma parte del objeto material de la física implica que las teorías psicológicas deben re-

[8] En beneficio de la exposición, supondré generalmente que las ciencias tratan sobre hechos, al menos en el sentido de que es la presencia de hechos lo que da veracidad a las leyes de una ciencia. Sin embargo, no hay nada que dependa de esta suposición.

ducirse a teorías físicas, y es este último principio el que complica las cosas. Yo pretendo evitar las complicaciones rechazando esa inferencia.

El reduccionismo es una opinión según la cual todas las ciencias especiales se reducen a la física. Sin embargo, el sentido de «reducirse a» es especial. Se puede describir de la siguiente manera[9].

Supongamos que la fórmula (1) es una ley de la ciencia especial *S*.

(1) $S_1x \rightarrow S_2y$

La fórmula (1) debería interpretarse algo parecido a «todos los hechos que consisten en que x sea S_1 dan lugar a hechos que consisten en que y sea S_2». Supongo que toda ciencia se individualiza por referencia a sus predicados característicos (véase nota 2, más arriba), y que por lo tanto, si S es una ciencia especial, «S_1» y «S_2» no son predicados de física básica. (También doy por supuesto que el «todos» que cuantifica las leyes de las ciencias especiales debe interpretarse con ciertas reservas. Estas leyes *no* son sin excepción. Volveré a ocuparme de esta cuestión más detalladamente.) Una condición necesaria y suficiente para la reducción de la fórmula (1) a una ley de física es que las fórmulas (2) y (3) sean leyes, y una condición necesaria y suficiente para

(2a) $S_1x \rightleftharpoons P_1x$
(2b) $S_2y \rightleftharpoons P_2y$
(3) $P_1x \rightarrow P_2y$

la reducción de S a física es que todas sus leyes se reduzcan de esta manera[10].

Se supone que «P_1» y «P_2» son predicados de física y que la fórmula (3) es una ley física. Las fórmulas del tipo de la fórmula (2) se denominan frecuentemente leyes «puente». Su rasgo característico es que contienen predicados de la ciencia reducida y de la reductora. Las leyes puente, como la fórmula (2), están por lo tanto en contraste con las leyes «propiamente dichas», como las fórmulas (1) y (3). La conclusión de las observaciones precedentes es que la reducción de una ciencia requiere que toda fórmula que aparezca como antecedente o consiguiente de una de sus leyes propiamente dichas debe aparecer como fórmula reducida en una de las leyes puente[11].

---

[9] La versión del reduccionismo de que me voy a ocupar es más fuerte de lo que mantienen muchos filósofos de la ciencia, lo cual es digno de tenerse en cuenta, ya que mi argumentación será precisamente que es demasiado fuerte. Sin embargo, creo que lo que voy a atacar es lo que muchas personas tienen presente cuando se refieren a la unidad de la ciencia, y sospecho (aunque no voy a tratar de demostrarlo) que muchas de las versiones liberalizadas del reduccionismo tienen el mismo defecto básico que lo que voy a considerar como forma clásica de tal doctrina.

[10] Existe la suposición implícita de que una ciencia *es* sencillamente la formulación de un conjunto de leyes. Creo que esta suposición es poco plausible, pero suele hacerse cuando se trata el tema de la unidad de la ciencia, y es neutral en lo que se refiere a la argumentación central de este capítulo.

[11] En algunas ocasiones me referiré a «el predicado que constituye el antecedente o consecuente de una ley». Esto constituye una forma resumida de hacer mención a «el predicado tal que el antecedente o consecuente de una ley consista en ese predicado, junto con sus variables ligadas y los cuantificadores que las unen». (Las funciones de verdad de los predicados elementales son en cuanto tales predicados en este sentido.)

Conviene hacer algunas indicaciones sobre el signo «→» de conexión. En primer lugar, cualesquiera que sean las propiedades de este elemento conectivo, todos están de acuerdo en que debe ser transitivo. Esto es importante porque se suele suponer que la reducción de alguna de las ciencias especiales se realiza a través de leyes puente que conectan sus predicados con los de las teorías reductoras intermedias. Así, se presupone que la psicología se reduce a física a través, por ejemplo, de la neurología, bioquímica y otras paradas locales. Lo que queremos señalar ahora es que esto no supone ninguna diferencia para la lógica de la situación con tal que se admita la transitividad de «→». Las leyes puente que conectan los predicados de S con los de S^* cumplirán con las constricciones de la reducción de S a física en la medida en que haya otras leyes puente que, directa o indirectamente, conecten los predicados de S^* con los predicados físicos.

Sin embargo, quedan abiertas cuestiones muy importantes sobre la interpretación de «→» en las leyes puente. Lo que está en juego en estas cuestiones es la medida en que se supone que el reduccionismo es una tesis fisicista.

Por de pronto, si en las leyes propiamente dichas interpretamos «→» como «dar lugar a» o «causar», tendremos que contar con otro conectivo para las leyes puente, pues dar lugar a y causar son probablemente asimétricos, mientras que las leyes puente expresan relaciones simétricas. Además, a no ser que las leyes puente tengan validez en virtud de la *identidad* de los hechos que cumplen con sus antecedentes y los que cumplen con sus consecuentes, el reduccionismo sólo nos asegurará una versión mitigada del fisicismo, y con ello no se conseguiría expresar el sesgo ontológico subyacente en el programa reduccionista.

Si las leyes puente no son afirmaciones de identidad, las fórmulas como la (2) lo más que pueden afirmar es que, por ley, la satisfacción por x de un predicado de P y la satisfacción por x de un predicado de S están en correlación causal. De aquí se desprende que es nomológicamente necesario que los predicados de S y de P se apliquen a las mismas cosas (es decir, que el predicado de S se aplique a un subconjunto de las cosas a que se aplican los predicados de P). Pero, evidentemente, esto es compatible con una ontología no fisicista, pues es compatible con la posibilidad de que el que x satisfaga S no tenga que ser un hecho físico. Según esta interpretación, la verdad del reduccionismo *no* garantiza la generalidad de la física frente a las ciencias especiales, pues hay algunos hechos (los que satisfacen a los predicados de S) que caen dentro de los dominios de una ciencia especial (S) pero no en el dominio de la física. (Podríamos pensar, por ejemplo, en una doctrina según la cual se considere que los predicados físicos y psicológicos se aplican a los organismos, pero en la que se niegue que el hecho que consiste en que un organismo satisfaga un predicado psicológico sea, en ningún sentido, un hecho físico. El resultado sería una especie de dualismo psicofísico de corte no cartesiano; un dualismo de hechos y/o propiedades en vez de sustancias).

Teniendo en cuenta estas consideraciones, muchos filósofos han afirmado que habría que suponer que las leyes puente como la fórmula (2) expresan identidades de hechos contingentes, de forma que la interpretación de la fórmula (2a) sería que «todo hecho que consiste en que x satisfaga S_1 es idéntico a otro hecho que consista en que x satisfaga P_1 y viceversa». Según esta lectura, la verdad del reduccionismo implicaría que todo hecho que cae dentro de una ley científica es un hecho físico, por lo

que expresaría simultáneamente el sesgo ontológico del reduccionismo y garantizaría la generalidad de la física frente a las ciencias especiales.

Si las leyes puente expresan identidades de hechos, y si todo hecho que se incluye dentro de las leyes propiamente dichas de una ciencia especial se incluye dentro de una ley puente, estamos ante el reduccionismo clásico, doctrina que implica la verdad de lo que podríamos llamar «fisicismo de hechos». Este fisicismo es sencillamente la afirmación de que todos los hechos de que hablan las ciencias son hechos físicos. Habría que hacer tres observaciones sobre el fisicismo de hechos.

En primer lugar, es menos fuerte que lo que se suele llamar «materialismo». El materialismo afirma que el fisicismo de hechos es cierto *y* que todo hecho cae dentro de las leyes de una u otra ciencia. Por consiguiente, se podría ser fisicista de hechos sin ser materialista, aunque no veo ninguna razón para tomarse esa molestia.

En segundo lugar, el fisicismo de hechos es menos fuerte que lo que podríamos llamar «fisicismo de propiedades», o doctrina que afirmaría, más o menos, que toda *propiedad* mencionada en las leyes de una ciencia es una propiedad física. El fisicismo de hechos no implica el fisicismo de propiedades, aunque sólo fuera porque la identidad contingente de un par de hechos no se supone que garantice la identidad de las propiedades de las que esos hechos son instancias particulares; ni siquiera cuando la identidad de los hechos es necesaria nomológicamente. Por otra parte, si un hecho es sencillamente la instanciación de una propiedad, el fisicismo de propiedades implica el fisicismo de hechos; dos hechos serán idénticos cuando consistan en la instanciación de la misma propiedad por el mismo individuo al mismo tiempo.

En tercer lugar, el fisicismo de hechos es menos fuerte que el reduccionismo. Como este punto constituye, en cierto sentido, el tema central del argumento siguiente, no lo desarrollaré aquí. Pero, como primera aproximación, el reduccionismo es la conjunción del fisicismo de hechos con la suposición de que hay predicados de clase natural en una física supuestamente terminada que corresponden a cada predicado de clase natural en una ciencia especial supuestamente terminada. Una de mis conclusiones será que no se puede deducir el reduccionismo a partir de la suposición de que es cierto el fisicismo de hechos. El reduccionismo es condición suficiente, pero no necesaria, del fisicismo de hechos.

Resumiendo: voy a suponer que el reduccionismo implica el fisicismo de hechos, pues si las leyes puente expresan identidades de hechos contingentes nomológicamente necesarias, la reducción de la psicología a la neurología supondría que todo hecho que consiste en la instanciación de una propiedad psicológica sea idéntico a otro hecho que consiste en la instanciación de una propiedad neurológica. Tanto el reduccionismo como el fisicismo de hechos presuponen la generalidad de la física, pues los dos afirman que todo hecho que cae dentro del universo del discurso de una ciencia especial caerá también dentro del universo del discurso de la física. Además, de ambas doctrinas se deduce la consecuencia de que toda predicción que se derive de las leyes de una ciencia especial (junto a una afirmación de las condiciones iniciales) se seguirá igualmente de una teoría que se componga únicamente de la física y las leyes puente (junto con la afirmación de las condiciones iniciales). Finalmente, en el reduccionismo y en el fisicismo de hechos se da por supuesto que la física es la *única* ciencia básica; *es decir,* que es la única ciencia que es general en los sentidos que acabamos de especificar.

Lo que ahora quiero demostrar es que el reduccionismo es una constricción demasiado fuerte para la unidad de la ciencia, pero que, para cualquier objetivo razonable, bastará con la doctrina menos fuerte.

Toda ciencia implica una taxonomía de los hechos dentro de su universo de discurso. En concreto, toda ciencia emplea un vocabulario descriptivo de los predicados teóricos y de observación, de manera que los fenómenos caen dentro de las leyes de la ciencia gracias a que cumplen con estos predicados. Evidentemente, no toda descripción verdadera de un hecho es una descripción con ese vocabulario. Por ejemplo, hay gran número de hechos que consisten en que ciertas cosas han sido transportadas a una distancia de menos de tres millas de la torre Eiffel. Sin embargo, estoy seguro de que no hay ninguna ciencia que contenga «es transportado a una distancia de menos de tres millas de la torre Eiffel» dentro de su vocabulario descriptivo. De la misma manera, doy por sentado que no hay ninguna ley natural que se aplique a los hechos en virtud de que constituyan un caso de la propiedad *es transportado a una distancia de menos de tres millas de la torre Eiffel* (aunque supongo que se puede pensar que hay una ley que se aplique a hechos en virtud de que constituyan un caso de una propiedad distinta pero coextensiva). Resumiendo, diré que la propiedad *es transportado...* no determina una clase (natural), y que los predicados que expresan esa propiedad no son predicados de clase (natural).

Si supiera lo que es una ley, y si creyera que las teorías científicas se componen únicamente de conjuntos de leyes, podría decir que «*P*» es un predicado de clase en relación a *S* si y sólo si *S* contiene leyes propiamente dichas del tipo «$P_x \rightarrow \ldots y$» o «$\ldots y \rightarrow P_x$»: en términos aproximados, los predicados de clase de una ciencia son aquellos cuyos términos son las variables ligadas en sus leyes propiamente dichas. Me siento inclinado a decir esto incluso en mi actual estado de ignorancia, aceptando la consecuencia de que con ello la oscura noción de clase tenga que depender de las igualmente oscuras nociones de *ley* y *teoría*. No podemos pisar suelo firme en este punto. Si no estamos de acuerdo en lo que es una clase, es probable que también estemos en desacuerdo sobre lo que es una ley, y por las mismas razones. No sé cómo se puede salir de este círculo, pero creo que se pueden decir algunas cosas interesantes sobre el círculo en que nos encontramos.

Por ejemplo, ahora podemos describir en qué sentido el reduccionismo es una interpretación demasiado fuerte de la doctrina de la unidad de la ciencia. Si el reduccionismo está en lo cierto, *toda* clase es una clase física o es coextensiva con ella. (Toda clase *es* una clase física si las afirmaciones puente expresan identidades de propiedad nomológicamente necesarias, y toda clase es coextensiva con una clase física si las afirmaciones puente expresan identidades de hechos nomológicamente necesarias.) Esto es algo que se sigue inmediatamente de la premisa reduccionista de que todo predicado que aparece como antecedente o consecuente de una ley de una ciencia especial debe aparecer como uno de los predicados reducidos de una ley puente, junto con la suposición de que los predicados de clase son aquellos cuyos términos son las variables ligadas de las leyes propiamente dichas. Si, resumiendo, una ley física está relacionada con cada ley de una ciencia especial en la forma en que la fórmula (3) está relacionada con la fórmula (1), en ese caso todo predicado de clase de una ciencia especial está relacionado con un predicado de clase de la física en la for-

ma en que la fórmula (2) relaciona «S_1» y «S_2» con «P_1» y «P_2» respectivamente.

Quiero ahora señalar algunas razones que llevan a la convicción de que esta consecuencia es intolerable. No tratan de ser razones «tumbativas»; no podrían serlo, dado que la cuestión de si el reduccionismo es demasiado fuerte constituye en último término una cuestión *empírica.* (Podría resultar que el mundo fuera de tal manera que toda clase se corresponda con una clase física, lo mismo que podría ocurrir que fuera de tal manera que la propiedad *es transportada a una distancia de menos de tres millas de la torre Eiffel* determine una clase en hidrodinámica, por ejemplo. Lo que ocurre es que, tal como están las cosas, parece muy poco probable que el mundo resulte ser de una de estas dos maneras.)

La razón por la que es improbable que toda clase corresponda a una clase física es precisamente que a) muchas veces se pueden hacer generalizaciones interesantes (por ejemplo, generalizaciones que resisten contrafactuales) sobre hechos cuyas descripciones físicas no tienen nada en común; b) muchas veces ocurre que el que las descripciones físicas de los hechos subsumidos por estas generalizaciones tengan o no algo en común es, en un sentido obvio, totalmente irrelevante para la verdad de las generalizaciones, o para su interés, o para su grado de confirmación, o para cualquiera de sus propiedades epistemológicamente importantes, y c) las ciencias especiales se dedican en gran parte a formular generalizaciones de esta clase.

Quiero suponer que estas observaciones son evidentes hasta el punto de que se autocertifican; saltan a la vista en el momento en que se adopta la actitud (aparentemente radical) de tomar la existencia de las ciencias especiales con toda seriedad. Supongamos, por ejemplo, que la «ley» de Gresham es cierta. (Si alguien tiene antipatía a la ley de Gresham, es probable que pueda servir con la misma perfección cualquier generalización verdadera que soporte contrafactuales de cualquier economía futura imaginable.) La ley de Gresham dice algo sobre lo que ocurre en los intercambios monetarios en ciertas condiciones. Estoy dispuesto a aceptar que la física es general *en el sentido de que implica que todo hecho que consiste en un intercambio monetario* (y por tanto todo hecho que caiga dentro de la ley de Gresham) *tiene una verdadera descripción en el vocabulario de la física y en virtud de lo cual cae dentro de las leyes de la física.* Pero una consideración superficial nos hace pensar que una descripción física que abarque todos estos hechos tiene que ser tremendamente disyuntiva. Algunos intercambios monetarios se hacen con cuentas de concha. Otros con billetes de dólar. Y en otros casos hay que firmar el propio nombre en un talón. ¿Cuáles son las probabilidades de que una disyunción de predicados físicos que cubra todos estos hechos (es decir, un predicado disyuntivo que pueda constituir la parte de la derecha de una ley puente de la forma «x es un intercambio monetario $\rightleftharpoons$ ...») exprese una clase física? En concreto, ¿cuáles son las probabilidades de que dicho predicado forme el antecedente o consecuente de una ley propiamente dicha de física? Lo importante es que los intercambios monetarios tienen cosas interesantes en común; la ley de Gresham, si es verdadera, dice qué es una de estas cosas interesantes. Pero lo que hay de interesante en los intercambios monetarios no son sus aspectos comunes dentro de la descripción *física.* Una clase como la del intercambio monetario *podría* resultar coextensiva con una clase física; pero si lo fuera, eso constituiría un accidente a escala cósmica.

En realidad, la situación del reduccionismo es todavía peor de lo que haría pensar

lo que venimos diciendo. El reduccionismo afirma no sólo que todas las clases son coextensivas con las clases físicas, sino que las coextensiones son nomológicamente necesarias: las leyes puente son *leyes*. De esta manera, si la ley de Gresham es cierta, se puede deducir que existe una ley (puente) de la naturaleza en virtud de la cual «*x* es un intercambio monetario $\rightleftharpoons$ *x* es *P*» es cierta para todos los valores de *x*, y *P* es un término que designa una clase física. Pero, indudablemente, no hay una ley así. Si la hubiera, *P* tendría que incluir no sólo todos los sistemas de intercambio monetario *existentes,* sino también todos los sistemas de intercambio monetario que *pudieran existir;* una ley debe valer con las contrafactuales. ¿Qué predicado físico es candidato a *P* en «*x* es un intercambio monetario nomológicamente posible si y sólo si P_x»?

Resumiendo: un econofísico inmortal podría encontrar, cuando hubiera terminado el espectáculo, un predicado de física que fuera, en términos aproximados, coextensivo con «es un intercambio monetario». Si la física es general —si los sesgos ontológicos del reduccionismo son ciertos— tiene que haber un predicado semejante. Pero, a) parafraseando una observación que el profesor Conald Davidson hizo en un contexto ligeramente diferente, lo único que podría convencernos de esta coextensividad total sería una enumeración total; b) parece que no habría ninguna posibilidad en absoluto de que el predicado físico utilizado para afirmar la coextensividad fuera un término de clase físico, y c) serían todavía menores las posibilidades de que la coextensión tuviera carácter legal (es decir, de que se aplicara no sólo al mundo nomológicamente posible que resultara ser real, sino a cualquier mundo nomológicamente posible)[12].

---

[12] Oppenheim y Putnam (1958) afirman que es probable que las ciencias sociales se *puedan* reducir a física, suponiendo que la reducción se realiza a través de la psicología (individual). Así, observan dichos autores, «en economía, se satisfacen suposiciones muy débiles, es posible representar la forma en que un individuo ordena sus elecciones por medio de una función de preferencia individual. De acuerdo con estas funciones, el economista trata de explicar los fenómenos de grupo, como el mercado, para explicar la conducta colectiva del consumidor, para solucionar los problemas de la economía del bienestar, etc.» (p. 17). Sin embargo, parece que no se han dado cuenta de que aunque pudieran realizarse estas explicaciones, no darían lugar a la clase de reducción *predicado-por-predicado* de la economía a la psicología que se exige en la explicación que los mismos Oppenheim y Putnam hacen de la unidad de la ciencia.

Supongamos que las leyes de la economía son válidas porque las personas tienen las actitudes, motivos, objetivos, necesidades, estrategias, etc., que tienen de hecho. Entonces, el hecho de que la economía sea tal como es se puede explicar haciendo referencia al hecho de que las personas son tal como son. Pero de eso no se deduce que los predicados característicos de la economía se puedan reducir a los predicados característicos de la psicología. Como las leyes puente implican bicondicionales, P_1 se reduce a P_2 únicamente si P_1 y P_2 son cuando menos coextensivos. Pero mientras los predicados característicos de la economía subsumen (por ejemplo) sistemas monetarios, los movimientos de efectivos, mercancías, recursos en mano de obra, cantidades de capital invertido, etc., los predicados característicos de la psicología subsumen estímulos, respuestas y estados mentales. Dado el sentido particular de la «reducción» de que estamos hablando, reducir la economía a psicología implicaría mucho más que demostrar que la conducta económica de los grupos está determinada por la psicología de los individuos que los constituyen. En concreto, supondría demostrar que nociones como las de *mercancía, recursos en mano de obra,* etc., se pueden reconstruir utilizando el vocabulario de estímulos, respuestas y estados mentales y que, además, los predicados que afectan a la reconstrucción expresan clases psicológicas (es decir, ocurren en las leyes propiamente dichas de la psicología). Creo que es justo decir que no hay absolutamente ninguna razón para suponer que es posible realizar estas reconstrucciones; a primera vista, tenemos todos los motivos para pensar que es imposible.

Supongo que la exposición precedente indica claramente que la economía no es reducible a la física en el sentido particular de la reducción que se da en las afirmaciones de la unidad de la ciencia. En este sentido, supongo yo, no hay nada de especial en relación con la economía; las razones por las que la economía no tiene muchas probabilidades de reducirse a física se dan en forma paralela en las que hacen pensar que no es probable que la psicología se reduzca a neurología.

Si la psicología es reducible a neurología, por cada predicado de clase psicológico existe un predicado de clase neurológico coextensivo con el anterior, y la generalización que afirma esta coextensión tiene carácter de ley. Evidentemente, hay muchos psicólogos que piensan algo parecido. Existen departamentos de psicobiología o de psicología y ciencia del cerebro en algunas universidades de todo el mundo cuya misma existencia es una apuesta institucionalizada en el sentido de que es posible encontrar estas coextensiones con carácter de ley. Sin embargo, como se ha resaltado frecuentemente en las discusiones actuales sobre el materialismo, existen motivos suficientes para no arriesgarse a hacer tales apuestas. No hay datos firmes, como no sean de una correspondencia muy vaga entre tipos de estados psicológicos y tipos de estados neurológicos, y es totalmente posible que el sistema nervioso de los organismos superiores consiga un determinado fin psicológico a través de una amplia variedad de medios neurológicos. Es también posible que determinadas estructuras neurológicas tengan relación con muchas funciones psicológicas diferentes en momentos distintos, en dependencia del carácter de las actividades en que interviene el organismo[13]. En cualquier caso, el intento de emparejar las estructuras neurológicas con las funciones psicológicas sólo puede aspirar a tener un éxito limitado. Psicólogos fisiológicos de la talla de Karl Lashley han mantenido esta opinión.

Lo que queremos decir en este momento es que el programa reduccionista *no* se puede defender en psicología basándose en motivos ontológicos. Aun cuando los hechos psicológicos (uno por uno) sean hechos neurológicos (uno por uno), no se deduce que los predicados de clase de la psicología sean coextensivos con los predicados de clase de cualquier otra disciplina (incluyendo la física). Es decir, la suposición de que todo hecho psicológico es un hecho físico no garantiza que la física (o, a fortiori, cualquier otra disciplina más general que la psicología) pueda proporcionar un vocabulario apropiado para las teorías psicológicas. Insisto en este punto porque estoy convencido de que el compromiso de todo-o-nada de muchos psicólogos fisiológicos con el programa reduccionista procede precisamente del hecho de haber confundido tal programa con el fisicismo (de hechos).

Lo que he estado poniendo en duda es que haya clases neurológicas coextensivas con clases psicológicas. Lo que parece cada vez más claro es que, aun cuando existan tales coextensiones, no pueden tener carácter de ley. Parece cada vez más probable que haya sistemas nomológicamente posibles que sean diferentes de los organismos (a saber, los autómatas) que satisfagan los predicados de clase de la psicología pero que no satisfagan ningún predicado neurológico en absoluto. Ahora bien, como ha

[13] Esto es lo que ocurriría si los organismos superiores resultan sorprendentemente análogos a los ordenadores de aplicación general. Estos aparatos no manifiestan a lo largo del tiempo ninguna correspondencia detallada estructura-función; por el contrario, la función que se realiza por medio de una determinada estructura puede variar de un instante a otro según el carácter del programa y de la computación realizada.

señalado Punam (1960a, b.), si existen estos sistemas, su número tiene que ser muy elevado, pues, en principio, se pueden hacer autómatas semejantes a partir prácticamente de todo. Si esta observación es correcta, no se puede tener ninguna esperanza fundada de que la clase de autómatas cuya psicología sea efectivamente idéntica a la de un organismo se pueda describir en predicados de clase *físicos* (aunque, naturalmente, si el fisicimo de hechos está en lo cierto, dicha clase puede ser recogida en uno u otro predicado físico). El resultado es que la formulación clásica de la unidad de la ciencia está a merced del progreso en el campo de la simulación con ordenador. Lógicamente, con esto lo único que se quiere decir es que tal formulación era demasiado fuerte. La unidad de la ciencia estaba destinada a ser una hipótesis empírica, falseable mediante posibles descubrimientos científicos. Pero nadie pensaba en que se iba a ver derrotada por Newell, Shaw y Simon.

Hasta ahora he expuesto que el reduccionismo psicológico (doctrina de que toda clase natural psicológica es, o es coextensiva con, una clase natural neurológica) no es equivalente a, ni puede deducirse de, el fisicismo de hechos (doctrina según la cual todo hecho psicológico es un hecho neurológico). Sin embargo, se puede argumentar que podrían considerarse estas doctrinas como equivalentes, dado que la única posible prueba que se podría tener a favor del fisicismo de hechos sería también una prueba a favor del reduccionismo: es decir, que esa prueba tendría que consistir en el descubrimiento de correlaciones psicofísicas de un tipo con otro.

Pero si nos detenemos a considerar esto por unos momentos, comprobaremos que esta argumentación no está bien desarrollada. Si las correlaciones psicofísicas de un tipo con otro constituyeran una prueba a favor del fisicismo de hechos, también lo harían las correlaciones de otras clases especificables.

Tenemos correlaciones de un tipo con otro allí donde por cada *n*-tuplo de hechos que sean de la misma clase psicológica se da un *n*-tuplo correlacionado de hechos que son de la misma clase neurológica[14]. Imaginemos un mundo en el que *no* se den estas correlaciones, en que lo que encontramos es que por cada *n*-tuplo de hechos psicológicos idénticos de tipo se da un *n*-tuplo correlacionado espaciotemporalmente de hechos neurológicos *distintos* de tipo. Es decir, todo hecho psicológico está emparejado con algún hecho neurológico, pero los hechos psicológicos de la misma clase están a veces emparejados con hechos neurológicos de diferentes clases. Lo que quiero destacar en este punto es que estos emparejamientos constituirían una base para el fisicismo de hechos en la misma medida que podrían hacerlo los emparejamientos tipo-con-tipo *con tal que seamos capaces de demostrar que los hechos neurológicos distintos de tipo emparejados con una clase determinada de hechos psicológicos son idénticos en relación a cualesquiera de las propiedades relacionadas con la identificación del tipo en psicología*. Supongamos, a efectos de explicación, que los hechos psicológicos se identifican en cuanto al tipo por relación a sus consecuencias conductuales[15]. Entonces, lo que se exige de todos los hechos neurológicos emparejados con una cla-

[14] Para eliminar los casos anómalos, suponemos que *n* es lo suficientemente grande como para dar lugar a correlaciones que sean significativas en sentido estadístico.

[15] Creo que no hay ninguna posibilidad de que esto sea cierto. Lo que sí es más probable es que la identificación del tipo en relación con los estados psicológicos se pueda realizar en relación con los «estados totales» de un autómata abstracto que reproduzca al organismo a que pertenecen los estados. Para una exposición más amplia, véase Block y Fodor (1972).

se de hechos psicológicos homogéneos en cuanto al tipo es únicamente que sean idénticos en relación con sus consecuencias conductuales. Por decirlo con pocas palabras, los hechos idénticos en cuanto al tipo no tienen, lógicamente, en común *todas* sus propiedades, y los hechos distintos en cuanto al tipo deben sin embargo ser idénticos en *algunas* de sus propiedades. La confirmación empírica del fisicismo de hechos no depende de que se demuestre que las réplicas neurológicas de los hechos psicológicos idénticos en cuanto al tipo son ellas mismas idénticas en cuanto al tipo. Lo que hay que demostrar es que son idénticas en relación a aquellas propiedades que determinan qué clase de hecho *psicológico* es un hecho determinado.

¿Podríamos tener pruebas de que un conjunto de hechos neurológicos, heterogéneos en los demás sentidos, tienen esas clases de propiedades en común? Por supuesto que sí. La teoría neurológica podría explicar por qué un *n*-tuplo de hechos neurológicamente distintos en cuanto al tipo son idénticos en sus consecuencias conductuales, o en relación con algunas de las otras propiedades relacionales, en número indefinido. Y si la ciencia neurológica no lo consiguiera, podría hacerlo alguna otra ciencia más básica que la neurología.

Lo que quiero decir con todo esto, una vez más, no es que las correlaciones entre estados psicológicos homogéneos en cuanto al tipo y los estados neurológicos heterogéneos en cuanto al tipo vayan a probar que el fisicismo de hechos es verdadero. Sólo quiero decir que estas correlaciones podrían darnos tantas razones para ser fisicistas de hechos como las correlaciones tipo-con-tipo. Si esto es correcto, los argumentos epistemológicos que pasan del fisicismo de hechos al reduccionismo no pueden dejar de ser erróneos.

Me parece (hablando en sentido muy general) que la interpretación clásica de la unidad de la ciencia ha cometido graves errores en la interpretación del *objetivo* de la reducción científica. El sentido de la reducción *no* es fundamentalmente encontrar un predicado de clase natural de la física que sea coextensivo con cada uno de los predicados de clase de una ciencia especial. Es, más bien, explicar los mecanismos físicos por los cuales los hechos se conforman a las leyes de las ciencias especiales. He estado intentando demostrar que no hay ninguna razón lógica o epistemológica en virtud de la cual el éxito en el segundo de estos proyectos implique el éxito en el primero, y que es probable que, *de hecho,* ambos se den separadamente siempre que los mecanismos físicos por los cuales los hechos se conforman a una ley de las ciencias especiales sean heterogéneos.

Supongo que lo que hemos tratado hasta ahora habrá servido para hacer ver que el reduccionismo es, probablemente, demasiado fuerte en cuanto interpretación de la unidad de la ciencia; por una parte, es incompatible con los resultados probables de las ciencias especiales, y, por la otra, es más de lo que necesitamos presuponer si lo que deseamos fundamentalmente, desde un punto de vista ontológico, es sencillamente ser buenos fisicistas de hechos. En las páginas que siguen, trataré de esbozar una versión liberalizada de la relación entre física y ciencias especiales que me parece tener exactamente la fuerza necesaria en estos sentidos. Luego presentaré un par de razones independientes que permiten pensar que quizá sea esta doctrina revisada la correcta.

En todo momento el problema ha sido que existe una posibilidad empírica de que

lo que corresponde a los predicados de clase de una ciencia reducida sea una disyunción heterogénea y asistemática de predicados en la ciencia reductora. No queremos que con esta posibilidad se prejuzgue la unidad de la ciencia. Supongamos, entonces, que accedemos a que las fórmulas puente sean de esta forma,

(4) $Sx \rightleftharpoons P_1x \vee P_2x \vee \ldots \vee P_nx$

donde $P_1 \vee P_2 \vee \ldots \vee P_n$ *no* es un predicado de clase en la ciencia reductora. Considero que esto equivale a admitir que al menos algunas «leyes puente» pueden, de hecho, no ser leyes, pues considero que una condición necesaria para que una generalización universal tenga carácter de ley es que los predicados que constituyen su antecedente y consecuente sean predicados de clase. De esta manera estoy suponiendo que, en relación con la unidad de la ciencia, es suficiente que todas las leyes de las ciencias especiales sean reducibles a la física mediante fórmulas puente que expresen verdaderas generalizaciones empíricas. Teniendo en cuenta que las fórmulas puente deben interpretarse como especies de afirmaciones de identidad, la fórmula (4) se interpretará algo parecido a «todo hecho que consiste en que x satisfaga S es idéntico a otro hecho que consista en que x satisfaga uno u otro de los predicados pertenecientes a la disyunción $P_1 \vee P_2 \vee \ldots \vee P_n$».

Ahora bien, en casos de reducción donde lo que corresponde a la fórmula (2) no es una ley, tampoco lo será lo que corresponde a la fórmula (3), y por la misma razón: a saber, los predicados que aparecen en el antecedente y consecuente no son, por hipótesis, predicados de clase. Lo que tengamos será algo que se parezca a la Figura I-1. Es decir, el antecedente y consecuente de la ley reducida estarán conectados con una disyunción de predicados en la ciencia reductora. Supongamos, por el momento, que la ley reducida no admite excepciones, es decir, que ningún hecho de S_1 satisface P'. Entonces, habrá leyes de la ciencia reductora que conecten la satisfacción de *cada* miembro de la disyunción asociada al antecedente de la ley reducida con la satisfacción de algún miembro de la disyunción asociada al consecuente de la ley

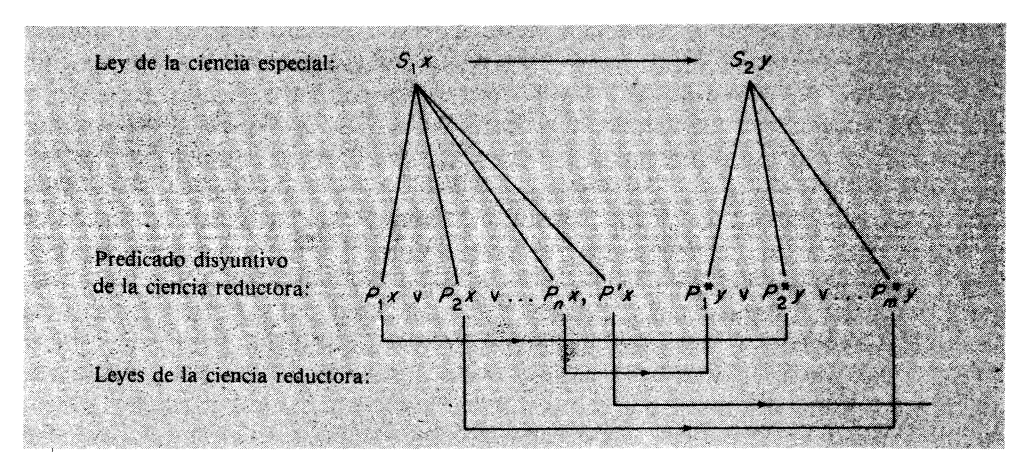


FIGURA I-1. *Representación esquemática de la relación propuesta entre la ciencia reducida y la ciencia reductora, en una explicación revisada de la unidad de la ciencia. Si algunos hechos* S_1 *son del tipo* P′, *serán excepción a la ley* $S_1x \rightarrow S_2y$. *Véase el texto.*

reducida. Es decir, si $S_1x \rightarrow S_2y$ no tiene excepciones, tiene que haber una ley propiamente dicha de la ciencia reductora que o afirme o implique que $P_1x \rightarrow P^*$ para algún P^*, y de la misma manera para P_2x y así sucesivamente hasta P_nx. Como tiene que haber leyes de este tipo, y como cada una de ellas es una ley «propiamente dicha» en el sentido en que venimos utilizando el término, se desprende que cada término de la disyunción P_1 v P_2 v ... v P_n es un predicado de clase, igual que cada término de la disyunción P^*_1 v P^*_2 v ... v P^*_n.

Y así llegamos al momento decisivo. Podría argumentarse que si cada término de la disyunción de los P está conectada mediante una ley a algún término de la disyunción de los P^*, se deduce que la fórmula (5) es en sí misma una ley.

(5) $P_1x \vee P_2x \vee \ldots \vee P_nx \rightarrow P^*_1y \vee P^*_2y \vee \ldots \vee P^*_ny$

La idea es que, según el esquema de la Figura I-1, $P_1x \rightarrow P^*_2y$, $P_2x \rightarrow P^*_my$, etc., y que el argumento que pasa de una premisa de la forma $(P \supset R)$ y $(Q \supset S)$ a una conclusión de la forma $(P \vee Q) \supset (R \vee S)$ es válido.

Lo que me inclino a decir al respecto es que lo anterior demuestra sencillamente que la expresión «es una ley que ____________» define un contexto funcional de no verdad (o, lo que vendría a ser lo mismo a este respecto, que no todas las funciones de verdad de los predicados de clase son, en cuanto tales, predicados de clase); en concreto, que no se puede pasar de: «es una ley que *P* da lugar a *R*» y «es una ley que *Q* da lugar a *S*» a «es una ley que *P* o *Q* da lugar a *R* o *S*». (Aunque, lógicamente, es válido el argumento que pasa de dichas premisas a «*P* o *Q* da lugar a *R* o *S*» *simpliciter*.) Creo, por ejemplo, que es una ley que la irradiación de las plantas verdes por la luz del sol produce la síntesis de los hidratos de carbono, y creo que es una ley que la fricción produce calor, pero no creo que sea una ley que (o la irradiación de las plantas por la luz del sol o la fricción) produzca (o la síntesis de los hidratos de carbono o calor). Por consiguiente, pongo en duda que «es o síntesis de hidratos de carbono o calor» se pueda considerar plausiblemente como un predicado de clase.

No es absolutamente obligatorio estar de acuerdo con todo esto, pero el negarlo supone pagar un precio. En concreto, si se admite toda la gama de argumentos con valor funcional de verdad dentro del contexto «es una ley que ____________», se renuncia a la posibilidad de identificar los predicados de clase de una ciencia con los que constituyen los antecedentes o consecuentes de sus leyes propiamente dichas. (Así la fórmula (5) sería una ley propiamente dicha de física que no cumple esa condición.) De esta manera se crea la necesidad de llegar a una nueva interpretación de la noción de clase, y no se me ocurre cuál podría ser.

La conclusión parece ser la siguiente: Si no exigimos que las fórmulas puente sean leyes, se sigue o bien que algunas de las generalizaciones a que se reducen las leyes de las ciencias especiales no son ellas mismas leyes, o bien que algunas leyes no son formulables en términos de clases. Cualquiera que sea la forma en que se considere la fórmula (5), lo importante es que la relación entre las ciencias propuesta por la Figura I-1 es más débil de lo que exige el reduccionismo al uso. En concreto, no supone una correspondencia entre los predicados de clase de la ciencia reducida y la ciencia reductora. Sin embargo sí que implica el fisicismo, si se da la misma suposición que hace que el reduccionismo al uso sea fisicista: es decir, que las afirmaciones puente

expresen identidades de hechos considerados individualmente. Pero éstas son precisamente las propiedades que queríamos que manifestara una explicación revisada de la unidad de la ciencia.

A continuación voy a presentar otras dos razones para admitir que esta interpretación de la unidad de la ciencia es correcta. En primer lugar, nos deja ver cómo las leyes de las ciencias especiales pueden tolerar razonablemente excepciones, y, en segundo lugar, nos hace comprender por qué existen ciencias especiales. Vayamos por partes.

Consideremos una vez más el modelo de reducción implícito en las fórmulas (2) y (3). Supongo que las leyes de la ciencia básica carecen absolutamente de excepciones, y supongo que es de dominio general que las leyes de las ciencias especiales sí las tienen. Pero en este momento tenemos que enfrentarnos con un dilema. Dado que «→» expresa una relación (o relaciones) que debe ser transitiva, la fórmula (1) sólo puede tener excepciones si las tienen las leyes puente. Pero si las leyes puente tienen excepciones, el reduccionismo pierde su mordiente ontológico, pues ya no podemos decir que todo hecho que consista en la satisfacción de un predicado-*S* consista en la satisfacción de un predicado-*P*. En resumen, dado el modelo reduccionista, no podemos suponer con coherencia que las leyes puente y las leyes básicas no tengan excepciones al mismo tiempo que suponemos que las leyes especiales sí las tienen. Pero no podemos aceptar la violación de las leyes puente a no ser que estemos dispuestos a invalidar la afirmación ontológica que constituye el punto central del programa reduccionista.

Podemos salir de esta situación (sin prejuicio para el modelo reduccionista) de una de dos maneras. Podemos renunciar a la afirmación de que las leyes especiales tienen excepciones o a la de que las leyes básicas no las tienen. Me atrevo a señalar que ambas alternativas son poco recomendables —la primera porque huye al enfrentarse con la realidad—. No existe ninguna posibilidad de que las generalizaciones verdaderas, que resisten a los contrafactuales, de, por ejemplo, la psicología, resulten ser válidas en absolutamente todas y cada una de las condiciones en que se satisfacen sus antecedentes. Aunque el espíritu es fuerte la carne es débil. Siempre habrá lapsus conductuales que sean fisiológicamente explicables pero que carezcan de interés desde el punto de vista de la teoría psicológica. Pero la segunda alternativa no es mucho mejor. Después de todo, puede ocurrir que las leyes de la ciencia básica tengan excepciones. Pero la cuestión está en si se quiere que la unidad de la ciencia dependa de la suposición de que las tienen.

No obstante, según la explicación resumida en la Figura I-1, todo funciona satisfactoriamente. Una condición nomológicamente suficiente para que se dé una excepción a $S_1x \rightarrow S_2y$ es que las afirmaciones puente identifiquen la presencia de la satisfacción de S_1 con la presencia de la satisfacción de un predicado-*P* que no esté vinculado en cuanto tal en forma de ley con la satisfacción de un predicado P^* (es decir, supongamos que S_1 está vinculado con P' de tal manera que no haya ninguna ley que conecte P' con ningún predicado que las afirmaciones puente asocien con S_2. En ese caso toda instanciación de S_1 que sea contingentemente idéntica a una instanciación de P' será un hecho que constituya una excepción a $S_1x \rightarrow S_2y$). Téngase en cuenta que, en este caso, no podemos suponer ninguna excepción a las leyes de la ciencia reductora pues, por hipótesis, la fórmula (5) no es una ley.

De hecho, estrictamente hablando, la fórmula (5) no ocupa ninguna posición en la reducción. Es sencillamente lo que se consigue cuando se cuantifica universalmente una fórmula cuyo antecedente es la disyunción física correspondiente a S_1 y cuyo consecuente es la disyunción física correspondiente a S_2. En cuanto tal, será verdadera cuando $S_1x \rightarrow S_2y$ no tenga excepciones y falsa en el caso contrario. Lo que realiza la función de expresar los mecanismos físicos mediante los cuales los *n*-tuplos de los hechos se conforman, o dejan de conformarse, a $S_1x \rightarrow S_2y$ no es la fórmula (5) sino las leyes que relacionan estrechamente los elementos de la disyunción $P^*_1 \vee P^*_2 \vee \ldots \vee P^*_n$ con elementos de la disyunción $P^*_1 \vee P^*_2 \vee \ldots \vee P^*_m$. Donde *existe* una ley que relacione un hecho que satisfaga uno de los términos de disyunción P con un hecho que satisfaga uno de los términos de la disyunción P^*, el par de hechos así relacionados se conformará a $S_1x \rightarrow S_2y$. Cuando un hecho que satisface un predicado P no está relacionado por ley con un hecho que satisface un predicado P^*, tal hecho constituirá una excepción a $S_1x \rightarrow S_2y$. Lo importante es que ninguna de las leyes que realizan estas conexiones estrechas necesita tener excepciones para que las tenga $S_1x \rightarrow S_2y$.

Formulemos esta exposición menos técnicamente: Si quisiéramos, podríamos *exigir* que las taxonomías de las ciencias especiales se correspondieran con la taxonomía de la física insistiendo en las distinciones entre las clases postuladas por las primeras siempre que resulte que corresponden a clases distintas de la última. Con esto *haríamos* que las leyes de las ciencias especiales no tuvieran excepciones si las leyes de la ciencia básica no las tuvieran. Pero también es probable que con ello perderían vigor las generalizaciones que queremos que expresen las ciencias especiales. (Si la economía tuviera que proponer tantas *clases* de sistemas monetarios como realizaciones físicas de sistemas monetarios existentes, las generalizaciones de la economía no tendrían excepciones. Pero, probablemente, esto no tendría mucho sentido, ya que no les quedaría a los economistas ninguna generalización más que hacer. La ley de Gresham, por ejemplo, tendría que ser formulada como una disyunción vasta y abierta sobre lo que ocurre en el sistema monetario$_1$ o sistema monetario$_n$ en condiciones que no admitirían una caracterización uniforme. No podríamos decir sin más lo que ocurre en los sistemas monetarios pues, por hipótesis, «es un sistema monetario» no corresponde a ningún predicado de clase de la física.)

De hecho, lo que hacemos es exactamente lo contrario. Admitimos que las generalizaciones de las ciencias especiales *tengan* excepciones, conservando así las clases a las que se aplican dichas generalizaciones. Pero como sabemos que las descripciones *físicas* de los miembros de estas clases pueden ser muy heterogéneas, y como sabemos también que los mecanismos físicos que vinculan la satisfacción de los antecedentes de estas generalizaciones con la satisfacción de sus consecuentes pueden ser igualmente diversos, esperamos al mismo tiempo que se den excepciones en las generalizaciones y que éstas se vean «justificadas» en el plano de la ciencia reductora. Este es uno de los aspectos en que se considera que la física es la ciencia base; sería mejor que las excepciones a *sus* generalizaciones (si es que hay alguna) fueran fortuitas, pues ya no se puede «ir más allá» para explicar el mecanismo por el que ocurren en ella las excepciones.

Con esto entramos en el tema de por qué existen ciencias especiales. El reduccionismo, como observamos anteriormente, huye a la vista de los hechos sobre la insti-

tución científica: la existencia de un amplio y entremezclado conjunto de disciplinas científicas especiales que parecen muchas veces avanzar sin prestar excesiva atención al requisito de que sus teorías deben resultar «a la larga» ser física. Quiero decir que la aceptación de este requisito desempeña un papel escaso o nulo en la validación práctica de las teorías. ¿Por qué ocurre esto? Probablemente, la respuesta reduccionista será *enteramente* epistemológica. Si cuando menos las partículas físicas no fueran tan pequeñas (si los cerebros estuvieran en el *exterior,* donde pudiéramos verlos), *en ese caso* haríamos física en vez de paleontología (neurología en vez de psicología, psicología en vez de economía, y así sucesivamente). Se puede dar una respuesta epistemológica: a saber, que incluso si los cerebros estuvieran en el exterior donde los pudiéramos mirar, no sabríamos, en la situación actual, qué habría que buscar. Carecemos del aparato teórico adecuado para la taxonomía psicológica de los hechos neurológicos.

Si resulta que la descomposición funcional del sistema nervioso se corresponde exactamente con su descomposición neurológica (anatómica, bioquímica, física), entonces sólo hay razones epistemológicas para estudiar la primera en vez de la segunda. Pero supongamos que no se da tal correspondencia. Supongamos que la organización funcional del sistema nervioso se cruza transversalmente con su organización neurológica. En ese caso, la existencia de la psicología depende no del hecho de que las neuronas sean desesperantemente pequeñas, sino más bien del hecho de que la neurología no cuenta con las clases requeridas por la psicología.

Estoy queriendo dar a entender, aproximadamente, que hay ciencias especiales no debido a la naturaleza de nuestra relación epistémica con el mundo, sino debido a la forma es que está integrado el mundo: no todas las clases (no todas las clases de cosas y hechos sobre los que hay que hacer importantes generalizaciones que resisten los contrafactuales) son, o se corresponden con, clases físicas. Una forma de exponer la opinión reduccionista clásica es decir que las cosas que pertenecen a clases físicas diferentes ipso facto no pueden tener en común ninguna de sus descripciones proyectables[16]: que si x e y difieren en aquellas descripciones en virtud de las cuales caen dentro de las leyes propiamente dichas de la física, deben diferir en aquellas descripciones en virtud de las cuales caen dentro de cualquier ley. Pero ¿por qué tenemos que creer que esto es así? Cualquier par de entidades, por muy diferentes que sean sus estructuras físicas, deben converger sin embargo en un número indefinido de sus propiedades. ¿Por qué no puede haber, entre esas propiedades convergentes, algunas cuyas interrelaciones en forma de ley confirmen las generalizaciones de las ciencias especiales? ¿Por qué, dicho en pocas palabras, los predicados de clase de las ciencias especiales no pueden *clasificar* de forma cruzada las clases naturales físicas?[17].

---

[16] Para la idea de proyectabilidad, véase Goodman (1965). Todos los predicados proyectables son predicados de clase, pero no, probablemente, al revés.

[17] Como, por cierto, es indudable que hacen los predicados de los lenguajes naturales. (Para una exposición más detallada, véase Chomsky, 1965).

Afirmar que las taxonomías empleadas por las ciencias especiales clasifican en forma cruzada clases físicas es negar que las ciencias especiales, junto con la física, constituyan una jerarquía. Negar que las ciencias constituyen una jerarquía es negar precisamente lo que yo digo que afirma la doctrina clásica de la unidad de la ciencia, en la medida en que afirma algo más que el fisicismo de hechos.

La física desarrolla la taxonomía de su objeto material que conviene mejor a sus objetivos: la formulación de leyes sin excepción que son básicas en los diferentes sentidos expuestos más arriba. Pero no es ésta la única taxonomía que se puede exigir si hay que atender a los objetivos de la ciencia en general: por ejemplo, si queremos expresar las generalizaciones verdaderas, resistentes a los contrafactuales, que se puedan hacer. Por eso hay ciencias especiales, con sus taxonomías especializadas, que se dedican a expresar algunas de estas generalizaciones. Para que la ciencia esté unificada, todas estas taxonomías deben aplicarse *a las mismas cosas*. Para que la física sea ciencia básica, sería mejor que cada una de estas cosas fuera una cosa física. Pero ya no hace falta que las taxonomías que emplean las ciencias especiales tengan que reducirse a la taxonomía de la física. No hace falta, y probablemente no es verdad.

Por mucho que lo intentan, a muchos filósofos les cuesta aceptar literalmente las cosas que dicen los no filósofos. Desde que dejó de estar de moda el verificacionismo, la mayoría de los filósofos han admitido —algunos incluso han insistido en ello— que las afirmaciones de los profanos son muchas veces verdaderas cuando se interpretan correctamente. Pero muchas veces la interpretación correcta no es fácil de encontrar y casi siempre resulta considerablemente diferente de lo que los profanos habían pensado. Así, durante algún tiempo, los filósofos decían que hablar sobre mesas y sillas es una forma elíptica y engañosa de referirse a la situación del propio campo visual y advertían que era probable que se vinieran abajo los fundamentos de la inferencia inductiva a no ser que los objetos físicos resultaran ser «constructos» hechos a partir de los fenómenos lógicamente homogéneos con las postimágenes. De hecho, sin embargo, se comprobó que el «hablar del objeto físico» merecía bastante menos análisis de lo que se había supuesto. Resultó que las mesas y sillas no eran postimágenes en absoluto, y la práctica de la inferencia inductiva se mantuvo en pie.

Pero aunque ahora el reduccionismo es algo que la misma epistemología propiamente dicha deplora, se mantiene en las discusiones filosóficas de los «constructos teóricos» de las ciencias. Las teorías psicológicas, en concreto, han producido a muchos filósofos la impresión de estar en condiciones de deshipostatización, y las advertencias de que la alternativa a la reducción es un escepticismo que sólo puede llevar a la ruina tienen un tono que resulta demasiado familiar. Sin embargo, el tema central de estas observaciones introductorias ha sido que los argumentos en favor de la reducción conductual o fisiológica de las teorías psicológicas no son, después de todo, muy convincentes. De hecho, lo que resulta verdaderamente interesante es el aceptar literalmente las teorías psicológicas y ver lo que nos dicen acerca de los procesos mentales. Es lo que me propongo hacer en las páginas siguientes.

Capítulo 1
PRIMERAS APROXIMACIONES

Soy vuestro único Presidente.

LYNDON B. JOHNSON

La línea principal de argumentación de este libro es como sigue:

1. Los únicos modelos psicológicos de los procesos cognitivos que parecen ser al menos remotamente plausibles representan a dichos procesos como computacionales.
2. La computación presupone un medio de computación: un sistema representacional.
3. Las teorías remotamente plausibles son preferibles a la ausencia de teorías.
4. Nos vemos así obligados provisionalmente a atribuir un sistema representacional a los organismos. «Obligados provisionalmente» significa: obligados en la medida en que atribuimos procesos cognitivos a los organismos y en la medida en que nos tomamos en serio las teorías de estos procesos actualmente existentes.
5. Constituye un *objetivo* de investigación razonable tratar de describir el sistema representacional al que nos vamos a mantener provisionalmente fieles.
6. Constituye una *estrategia* de investigación razonable tratar de deducir esta descripción a partir de los detalles de las teorías psicológicas que tengan más posibilidades de resultar verdaderas.
7. Puede que de hecho esta estrategia dé resultado: es posible presentar inferencias que sigan las líneas del punto 6 y que, si no precisamente apodícticas, tienen al menos a primera vista un aire de plausibilidad.

La categoría epistémica de estos puntos es bastante variada. Yo considero, por ejemplo, que el número 3 es una verdad evidente en sí misma y por lo tanto no exige ninguna justificación, como no sea hacer una llamada a la razón. Opino que el punto 4 se deduce de los puntos 1-3. Los puntos 5-7, por el contrario, tienen que justificarse *en la práctica*. Lo que se debe demostrar es que, de hecho, es productivo realizar una investigación psicológica siguiendo las líneas en ellos recomendadas. Los capítu-

los finales del libro están orientados en gran parte a demostrar precisamente esto. Por eso, según vayamos progresando, la discusión se verá más estrechamente vinculada a los descubrimientos empíricos y a sus interpretaciones.

Sin embargo, este capítulo se refiere principalmente a los puntos 1 y 2. Lo que trataré de demostrar es que, prescindiendo de las propias suposiciones sobre los *detalles* de las teorías psicológicas de la cognición, su estructura general presupone procesos computacionales subyacentes y un sistema representacional en que se realizan tales procesos. Con mucha frecuencia son hechos muy conocidos los que, en primera instancia, determinan los propios modelos de la vida mental, y este capítulo constituye principalmente una meditación sobre algunos de ellos. En breves palabras, trataré de algunas clases de teorías que, en mi opinión, la mayoría de los psicólogos cognitivos aceptarían en rasgos generales, por mucho que puedan estar en desacuerdo con sus aspectos concretos. Quiero hacer ver cómo, en todos y en cada uno de los casos, estas teorías presuponen la existencia y explotación de un sistema representacional de cierta complejidad en que se llevan a cabo los procesos mentales. Comienzo con las teorías de la decisión.

Doy por supuesto que es evidente en sí mismo que los organismos creen muchas veces que la conducta que producen es conducta de una clase determinada y que ello constituye con frecuencia parte de la explicación de la forma en que se comporta un organismo para atenerse a las creencias que tiene sobre la forma de conducta que produce[1]. Supuesto esto, el modelo siguiente se presenta como abrumadoramente plausible en cuanto explicación de cómo se decide al menos parte de la conducta.

8. El agente se encuentra en una determinada situación (*S*).
9. El agente cree que en *S* se le ofrece un determinado conjunto de opciones conductuales (B_1, B_2, ... B_n); es decir, dado *S*, B_1 a B_n son las cosas que el agente cree que puede hacer.
10. Se prevén las consecuencias probables que se seguirían de realizar cada una de las opciones que van de B_1 a B_n; es decir, el agente calcula un conjunto de hipótesis cuya forma aproximada sería «si se realiza B_1 en *S*, entonces, con cierta probabilidad, se seguiría C_i. Cuáles son las hipótesis que se consideran y las probabilidades que se asignan es algo que depende, naturalmente, de lo que el organismo sabe o cree acerca de situaciones como *S*. (Dependerá también de otras variables que son, desde el punto de vista del modelo presente, meras interferencias: presión temporal, cantidad de espacio de cálculo de que dispone el organismo, etc.).

[1] No estoy indicando que se trate, en sentido técnico, de una verdad *necesaria*. Pero sí soy de la opinión de que es la clase de proposición que sería absurdo tratar de confirmar (o rebatir) haciendo experimentos. Es posible (aunque muy difícil) imaginar una situación en que fuera razonable abandonar la práctica de apelar a las creencias de un organismo en los intentos de explicar su conducta: bien porque se había comprobado que tales apelaciones eran internamente incoherentes o porque se había observado que se conseguían mejores explicaciones con un aparato teórico distinto. Sin embargo, tal eomo están las cosas, no se ha demostrado tal incoherencia (a pesar de las publicaciones operacionalistas en contra) y nadie tiene la menor idea de cómo podría ser una opción teórica alternativa (a pesar de las publicaciones conductistas en contra). Un principio metodológico, al que me mantendré escrupulosamente fiel en las páginas siguientes, es que si no se tiene otra alternativa que suponer que *P*, entonces no queda otra alternativa que suponer que *P*.

11. Se atribuye un orden de preferencia a las consecuencias.
12. La elección de conducta realizada por el organismo está determinada en función de las preferencias y probabilidades asignadas.

Dos advertencias. En primer lugar, esto no es una teoría sino un esquema de teoría. No se darán predicciones sobre lo que optarán por hacer los organismos concretos en las ocasiones concretas hasta que se atribuyan valores a las variables; por ejemplo, hasta que se sepa cómo se describe *S*, qué opciones conductuales se consideran, a qué consecuencias se cree que lleva la explotación de las opciones, qué orden de preferencia atribuye el organismo a estas consecuencias y qué relación probabilidad-preferibilidad acepta el organismo. Esto quiere decir que, igual que en otros contextos, una teoría respetable de la forma en que se comporta un organismo presupone una amplia información sobre lo que sabe y valora el organismo. Los puntos 8-12 no tratan de ofrecer esta teoría, sino únicamente de determinar algunas de las variables en cuyos términos habría que articularla.

En segundo lugar, es evidente que el modelo está sumamente idealizado. No siempre, en una situación determinada, contemplamos todas (o, incluso, ninguna de) las opciones conductuales que creemos tener en nuestra mano. Así como tampoco evaluamos siempre nuestras opciones a la luz de lo que consideramos que son sus consecuencias probables. (Los existencialistas, tengo entendido, insisten en no hacerlo nunca). Pero estas clases de alejamiento de los hechos no impugnan el modelo. Lo más que demuestran es que las conductas que producimos no están siempre en correspondencia racional con las creencias que mantenemos. Sin embargo, en relación con mi desarrollo basta con que algunos agentes sean racionales en alguna medida y durante algún tiempo, y que cuando lo sean, y en la medida en que lo sean, los procesos como los mencionados en los puntos 8-12 hagan de mediadores en la relación existente entre lo que el agente cree y lo que hace[2].

En la medida en que aceptamos que este modelo se aplica en algún caso determinado, aceptamos también las clases de explicación que el mismo autoriza. Por ejemplo, dado este modelo, podemos explicar el hecho de que el organismo *a* produjo la conducta *B* haciendo ver:

13. Que *a* creyó que estaba en la situación *S*.
14. Que *a* creía que produciendo una conducta del tipo B_i en *S* daría lugar probablemente a la consecuencia C_i.

---

[2] Naturalmente, no es condición suficiente para la racionalidad de la conducta que se impliquen en su producción procesos como los puntos 8-12. Por ejemplo, las conductas que tienen esa mediación serán generalmente *ir*racionales si las creencias implicadas en el punto 10 son supersticiosas, o si las preferencias implicadas en el punto 11 son perversas, o si las computaciones implicadas en los puntos 9-12 son gravemente erróneas. Tampoco, al menos en mi opinión, proponen los puntos 8-12 condiciones *lógicamente* necesarias de la racionalidad de la conducta. Volviendo a la expresión de la introducción, la historia conceptual de qué es lo que hace que la conducta sea racional, exige probablemente una cierta clase de correspondencia entre conducta y creencia, pero no se ocupa del carácter de los procesos mediante los que se realiza tal correspondencia; es lógicamente posible, supongo, que los ángeles sean racionales por reflejo. Lo que se afirma de los puntos 8-12 es sencillamente que ellos —o algo que tenga un parecido razonable con ellos— son *empíricamente* necesarios para producir una correspondencia racional entre las creencias y las conductas de las criaturas sublunares. Hay una forma concisa de indicar todo esto diciendo que los puntos 8-12 proponen una (esquemática) teoría psicológica.

15. Que C_i era una (o la) consecuencia altamente valorada por *a*.
16. Que *a* creía y pretendía que *B* fuera una conducta del tipo B_i.

Lo que hay que tener en cuenta es que en este modelo de explicación se implica que los agentes consideran algunas veces que su conducta es conducta de una clase determinada; en el caso que nos ocupa, parte de la explicación de la conducta de *a* es que él creía que era de la clase B_i, pues es a la conducta de esa clase a la que se vinculan las consecuencias tan apreciadas. Dicho en pocas palabras, la explicación no llega a *ser* una explicación (completa) de la conducta de *a* a no ser que la conducta fuera B_i y que *a* creyera que lo era.

Naturalmente, los puntos 13-16 podrían *contribuir* a una explicación de la conducta aun en el caso de que *no* se produzca *B* y de que el agente *no* considere que la conducta que de hecho ha producido es conducta del tipo B_i. «Will nobody pat my hiccup?», grita el epónimo reverendo Spooner. Suponemos que lo que trata de llegar a ser B_i es una descripción estructural del tipo de frase «Will nobody pick my hat up?» y que la disparidad entre la conducta producida y una muestra de ese tipo se puede atribuir a lo que las emisoras llaman una avería técnica momentánea*. En estos casos nuestra confianza en el hecho de que sabemos cuál era la conducta pretendida por el agente depende de tres convicciones:

17. Que los items 14 y 15 son verdaderos en la sustitución propuesta para B_i.
18. Que los items 14 y 15 serían falsos si en cambio tuviéramos que introducir una descripción del tipo de la cual la conducta observada era en realidad una muestra. (En el ejemplo que nos ocupa, se supone plausiblemente que Spooner no habría conseguido ningún beneficio positivo con la producción de una muestra del tipo «Will nobody pat mi hiccup?»; no se concibe ninguna razón que justifique que quiera decir *eso*.
19. Que es plausible hacer hipótesis sobre los mecanismos cuyas operaciones explicarían los aspectos en que difieren las conductas observada y la pretendida. (En el caso presente, mecanismos de metátesis)

Es claro que si las explicaciones «psicodinámicas» de la conducta son ciertas, los mecanismos considerados por el punto 19 pueden ser de una complejidad prácticamente impenetrable. En cualquier caso, lo que quiero destacar ahora es que la aplicabilidad de un esquema explicativo como los puntos 8-12 puede estar íntimamente presupuesta no sólo en las explicaciones de la conducta observada, sino también en las atribuciones de intenciones conductuales frustradas.

Estoy presentando estas observaciones evidentes porque creo que sus consecuencias inmediatas son de gran importancia para la construcción de las teorías cognitivas en general: es decir, que esta forma de explicación sólo puede progresar si supone-

* Las dos preguntas contienen un juego de palabras que no se puede traducir al castellano. Tampoco es importante para el desarrollo de la argumentación. Se trata de dos frases que tienen forma gráfica y fonética semejante pero que significan cosas completamente distintas. Es lo que en la «jerga» especializada recibe el nombre de «spoonerismos». (Algo parecido a «confundir la gimnasia con la magnesia»). La primera significaría: «¿es que nadie me va a dar golpes en la espalda para quitarme el hipo?», y la segunda: «¿es que nadie me va a coger el sombrero?». (N. del T.).

mos que los agentes tienen medios para representar sus conductas ante sí mismos; en realidad, medios para representar sus conductas en cuanto dotadas de ciertas propiedades y carentes de otras. En el caso presente, es esencial para la explicación que el agente pretenda y crea que la conducta que ha producido es una conducta de una determinada clase (es decir, de la clase asociada con las consecuencias relativamente muy valoradas que se dan en *S*) y no de otra clase (es decir, no de clase asociada con consecuencias relativamente poco valoradas que se dan en *S*). Si se renuncia a ello, renunciamos a la posibilidad de explicar la conducta del agente haciendo referencia a sus creencias y preferencias.

La conclusión a que quiero llegar es que ciertas clases de modelos muy importantes de explicación psicológica presuponen que el organismo que produce la conducta tiene a su disposición alguna forma de sistema representacional. En beneficio de la exposición, he subrayado la importancia de la representación que el organismo hace de su propia conducta al explicar las acciones consideradas. Pero, una vez hecha, parece que se trata de una afirmación omnipresente. Por ejemplo, en el modelo estaba implícito que el organismo dispone de medios para representar no sólo sus opiniones conductuales sino también: la consecuencia probable de actuar sobre tales opciones, una determinada ordenación por orden de preferencia de las consecuencias y, naturalmente, la situación original en que se encuentra. Utilizar este tipo de modelo es, por lo tanto, presuponer que el agente tiene acceso a un sistema representacional de riqueza muy considerable. Según el modelo, decidir es un proceso computacional; el acto que realiza el agente es consecuencia de computaciones definidas sobre las representaciones de las posibles acciones. Si no hay representaciones, no hay computaciones. Si no hay computaciones, no hay modelo.

Podría haber dicho también que el modelo presupone un lenguaje. Si hurgamos un poco, comprobaremos que el sistema representacional presupuesto por los puntos 8-12 tiene que compartir toda una serie de rasgos característicos con los lenguajes reales. Es este un tema al que volveré con considerable extensión en los capítulos 2 y 3. De momento nos contentaremos con señalar dos de las propiedades que este sistema de representaciones debe tener en común con los lenguajes propiamente dichos (por ejemplo, con los lenguajes naturales).

En primer lugar, el sistema debe contar con un número infinito de representaciones distintas. El argumento en este caso es análogo al que se da en favor del carácter no-finito de los lenguajes naturales: igual que, en este último caso, no hay un límite por arriba en la complejidad de una frase que se pueda utilizar para hacer una afirmación, en el primer caso no hay ningún límite por arriba en la complejidad de la representación que se pueda necesitar para especificar las opciones de conducta que están a disposición del agente, o la situación en que se encuentra, o las consecuencias de actuar en un sentido o en otro.

Naturalmente, con esto no se intenta demostrar que las posibilidades *prácticas* sean *literalmente* infinitas. De la misma manera que existe la-frase-más-larga-que-nadie-puede-emitir, tiene que haber la-más-compleja-situación-en-que-nadie-pueda-actuar. La capacidad infinita del sistema representacional es, por lo tanto, una idealización, pero no es una idealización *arbitraria*. En ambos casos, lo esencial es la capacidad del organismo para hacer frente a estimulaciones *nuevas*. Así, inferimos la productividad de los lenguajes naturales a partir de la capacidad del hablante/oyente pa-

ra producir/comprender oraciones en las que no ha recibido ningún adiestramiento específico. Es un argumento exactamente idéntico el que sirve para inferir la productividad del sistema representacional interno a partir de la capacidad del agente para calcular las opciones de conducta apropiadas a un tipo de situación con que nunca se ha encontrado anteriormente.

Pero no es la productividad la única propiedad importante común a los lenguajes naturales y al sistema de representación que se utilice para decidir qué hacer. Es evidente, por ejemplo, que la idea de que el agente puede representarse a sí mismo aspectos sobresalientes de las situaciones en que se encuentra presupone que propiedades semánticas tan familiares como la verdad y la referencia se manifiesten en fórmulas en el sistema representacional[3]. Hemos estado suponiendo que, por debajo de la capacidad de acción razonada, tiene que haber una capacidad para la descripción de las situaciones reales y posibles. Pero las nociones de descripción, verdad y referencia son inseparables: Más o menos, «*D*» describe a qué se refiere «*a*» si y sólo si («*Da*» es verdad si y sólo si *a* es *D*).

Una línea semejante de pensamiento demuestra que el sistema representacional deberá disponer de mecanismos para expresar las propiedades intensionales. En concreto, la acción deliberada presupone decisiones entre resultados posibles (pero) no-reales. Así, el sistema representacional utilizado para los cálculos debe distinguir entre situaciones posibles, no-reales. Aquí no voy a tratar de considerar el tema de si habría que hacer esto determinando los órdenes de preferencia respecto a las proposiciones (como indicarían los tratamientos tradicionales de la intensionalidad) o respecto a los mundos posibles (según los enfoques semánticos basados en la teoría de los modelos). Lo que quiero resaltar es sencillamente que tiene que haber *algún* mecanismo semejante a disposición del sistema representacional, y por razones paralelas a las que nos llevan a pensar que los lenguajes naturales pueden contar con alguno de estos mecanismos.

Hasta ahora he dado por sentado a lo largo de esta exposición que toda persona razonable aceptará que para una psicología de la elección es esencial algo parecido a los puntos 8-12; lo que he venido haciendo ha sido sencillamente desarrollar algunas de las implicaciones de tal presupuesto. Pero, evidentemente, el presupuesto no es cierto. Los conductistas, por ejemplo, no aceptan que decidir es un proceso computacional, por lo que las explicaciones conductistas de la acción no necesitan dar por sentado un sistema de representaciones internas. No tengo intención de considerar la cuestión de lo adecuado de estas explicaciones; me parece un tema muerto. Baste con decir que, a la luz de nuestra exposición, se puede profundizar en algunas de las críticas habituales.

Muchas veces se suele reprochar a los conductistas que buscan una reducción, a primera vista implausible, de las acciones calculadas a hábitos. Lo que se pretende decir con esta crítica es que, en la medida en que las acciones se consideran simple-

---

[3] Utilizo el término «fórmulas» sin prejuicio de cuáles puedan ser, de hecho, los vehículos de representación interna. En este momento de la exposición dejamos abierta la cuestión de si se trata de imágenes, o de las señales de un semáforo, o de frases en japonés. Gran parte del contenido de los siguientes capítulos se centra en lo que se sabe sobre el carácter de las representaciones internas y lo que se puede deducir al respecto a partir de lo que se sabe sobre otras cosas.

mente como respuestas adiestradas a los inputs ambientales, la productividad de la conducta se vuelve ininteligible. (Para un desarrollo más amplio, véase Chomsky, 1959). Pero no es éste el único fallo de quienes interpretan las conductas calculadas como una especie de respuestas condicionadas. Lo que todo el mundo sabe, aunque la metodología conductista se niegue a admitirlo, es que al menos algunas acciones son elecciones de entre una gama de opciones contempladas por el agente. El conductista no puede admitir esto porque se ve precisado a describir las acciones como resultados de causas ambientales. Como sólo los estados de cosas *reales* pueden ser causas, entre los determinantes de una respuesta no puede estar la-posibilidad-de-que-*P*. Pero, sin embargo, tampoco la *contemplación* de las posibilidades por el sujeto, pues, aunque ésta se pueda considerar como un hecho real en cualquier ontología racional, no es un hecho *ambiental* en el sentido particular que esa idea tiene para el conductista. De cualquier forma que se considere, el conductista está metodológicamente obligado a negar lo que parecería ser evidente en sí mismo: que algunas veces actuamos como actuamos porque nos parece la mejor manera de actuar teniendo en cuenta las opciones que consideramos. En resumen, el conductista nos obliga a opinar que las conductas deliberadas son respuestas a inputs reales, cuando lo que queremos hacer es considerarlas como respuestas a resultados posibles.

Por el contrario, una de las grandes ventajas de las teorías computacionales de la acción es que nos permiten reconocer lo que todos sabemos: que decidir lo que vamos a hacer implica muchas veces considerar lo que podría resultar de nuestra acción. Suponer un sistema representacional que pueda distinguir entre (es decir, atribuir representaciones diferentes a) los distintos estados de cosas es precisamente permitirse a uno mismo considerar la conducta que se produce realmente como una elección de entre aquellas opciones que el agente considera «abiertas». Vale la pena resaltar el hecho de que las publicaciones conductistas no ofrecen ninguna base en que apoyarse para rechazar esta interpretación tan plausible, a no ser la afirmación reiterada de que es, de alguna manera, «acientífica». Sin embargo, por lo que yo puedo ver, esto equivale únicamente a la observación (correcta) de que no se puede al mismo tiempo decir lo que es plausible decir sobre las acciones y adoptar una metodología conductista. Tanto peor para esta metodología.

Ya habrá imaginado el lector que lo que estoy proponiendo hacer es resucitar la idea tradicional de que hay un «lenguaje del pensamiento» y que un aspecto importante de lo que debe hacer una teoría de la mente es describir ese lenguaje. Se trata de un punto de vista con el que, me parece, gran parte de las actuales obras psicológicas sobre la cognición tienen una relación curiosa y ligeramente esquizoide. Por una parte, parece estar implícito en casi todas las clases de explicación que aceptan los psicólogos cognitivos pues, como he señalado anteriormente, la mayoría de estas explicaciones consideran a la conducta como resultado de la computación, y la computación presupone un medio en que realizarse. Pero, por otra parte, la admisión de este medio sólo se hace de forma explícita en casos relativamente raros, y la apremiante pregunta que obliga a plantearse —qué propiedades tiene el sistema de representaciones internas— sólo se convierte en objeto de investigación en casos ocasionales.

Yo me propongo, según vayamos avanzando, considerar diversos tipos de pruebas que pueden tener relación con la respuesta que se dé a dicha pregunta. Sin em-

bargo, antes de hacerlo quiero examinar otras dos líneas de argumentación que parecen conducir, con un cierto aire de inevitabilidad, a postular un lenguaje del pensamiento como precondición para todo tipo de construcción de teorías dentro de la psicología cognitiva. Mi punto de vista es que no es sólo la acción deliberada, sino también el aprendizaje y la percepción, lo que se debe interpretar como basado en procesos computacionales; y, una vez más, no hay computación sin representación.

Consideremos en primer lugar el fenómeno que los psicólogos denominan a veces «aprendizaje de conceptos». Quiero centrar la atención en el aprendizaje de conceptos no sólo porque constituye una ilustración útil de nuestra tesis principal (los procesos cognitivos son procesos computacionales y por lo tanto presuponen un sistema representacional) sino también porque el análisis del aprendizaje de conceptos repercute en una variedad de temas que aparecerán en los posteriores capítulos.

Para empezar, se puede decir que el aprendizaje de conceptos es uno de esos procesos en que, como consecuencia de sus experiencias, se produce un cambio en lo que sabe el organismo; en especial, como consecuencia de sus interacciones con el entorno. Pero, evidentemente, no *todos y cada uno* de los casos de una alteración en el conocimiento producida por el entorno se debe considerar como aprendizaje; *a fortiori,* no todos estos casos figuran como aprendizaje de *conceptos.* Así, por ejemplo, la afasia se induce muchas veces por medio del entorno, pero el volverse afásico no es una experiencia de aprendizaje. De la misma manera, si pudiéramos de alguna forma inducir el conocimiento del latín haciendo tomar unas píldoras azules, supongo que esto sería adquirir el latín sin aprenderlo. De la misma manera, la «impronta» (véase Thorpe, 1963) altera lo que sabe el organismo como consecuencia de sus experiencias, pero sólo es marginalmente un proceso de aprendizaje, si es que se puede decir en algún sentido que sea aprendizaje. En el mejor de los casos, una teoría general del aprendizaje de conceptos *no* es una teoría general de cómo afecta la experiencia al conocimiento.

Hay, además, clases de *aprendizaje* que no son probablemente clases de aprendizaje de conceptos[4]. El aprendizaje de memoria es un ejemplo plausible (por ejemplo, el aprendizaje de una lista de sílabas carentes de sentido. Sin embargo, véase Young, 1968). También lo es lo que podríamos denominar «aprendizaje sensorial» (aprender cómo sabe un filete, aprender cómo suena una nota en un oboe, y así sucesivamente). En términos muy aproximados, y con la intención exclusiva de determinar al área de que nos ocupamos, podríamos decir que lo que distingue el aprendizaje de memoria y el aprendizaje sensorial del aprendizaje de conceptos es que, en los primeros casos, lo que se *recuerda de* una experiencia agota lo que se *aprende de* esa experiencia. En cambio, el aprendizaje de conceptos va de alguna manera «más allá» de los datos experienciales. ¿Pero qué significa eso?

Creo que lo que tienen en común las situaciones de aprendizaje de conceptos es lo siguiente: Las experiencias que ocasionan el aprendizaje en tales situaciones (según

---

[4] Considero que esta cuestión es de carácter empírico; el que sea verdad o no, depende de lo que ocurra, de hecho, en los distintos procesos de aprendizaje. *Podría* ocurrir que el mecanismo de aprendizaje de conceptos fuera el mecanismo de aprendizaje general, pero sería una sorpresa si así fuera y yo quiero señalar explícitamente que no admito la suposición de que lo es. Necesitamos mucho —y no la tenemos— una taxonomía de las clases de aprendizaje que sea empíricamente defendible.

sus descripciones teóricamente relevantes) están en *relación de confirmación* con lo que se aprende (según *su* descripción teóricamente relevante). Podríamos abreviar diciendo que el aprendizaje de conceptos es esencialmente un proceso de formación y de confirmación de hipótesis[5]. La mejor forma de ver que esto es así es considerar el paradigma experimental en términos del cual el «constructo» de aprendizaje de conceptos viene a ser, como uno solía decir, «definido operacionalmente».

En la situación experimental característica, el sujeto (humano o infrahumano) se enfrenta con la tarea de determinar las condiciones del entorno en que resulta apropiada una respuesta que se le indica, y el aprendizaje se manifiesta en la tendencia cada vez mayor de *S*, al pasar el tiempo o los ensayos, a producir las respuestas indicadas cuando, y sólo cuando, se consiguen esas condiciones. La lógica del paradigma experimental exige, en primer lugar, que haya una «señal de error» (por ejemplo, reforzamiento o castigo o ambas cosas) que indique si se ha realizado adecuadamente la respuesta indicada y, en segundo lugar, que haya una «propiedad criterio» de los estímulos manejados experimentalmente de forma que el carácter de la señal de error esté en función de la presencia de la respuesta indicada junto con la presencia o ausencia de dicha propiedad. Así, en un experimento sencillo de esta clase, se podría indicar a *S* que clasificara las tarjetas-estímulos en montones, donde las figuras de las tarjetas contengan cualquier combinación de las propiedades rojo y negro con cuadrado y circular, pero donde la única clasificación correcta (por ejemplo, recompensada) es la que agrupa los círculos rojos con los cuadrados negros. En este caso, la «respuesta indicada» es hacer una clasificación en que se dé con el montón positivo y la «propiedad criterio» es *círculo rojo o cuadrado negro*.

Es posible utilizar este tipo de disposición experimental para estudiar la tasa de aprendizaje en función de variables muy diversas: por ejemplo, el tipo de la propiedad criterio; el tipo de la señal de error; la capacidad de *S* para referir cuál es la propiedad de criterio en que se basa su tarea; el tipo de relación (temporal, estadística, etcétera) entre las ocurrencias de la señal de error y las instanciaciones de la propiedad criterio; el tipo de la población a la que pertenece el sujeto (edad, especie, inteligencia, motivación, o cualquier otro aspecto), y así sucesivamente. Gran parte de la psicología experimental de aprendizaje se viene ocupando desde hace treinta años de los enormes cambios que se producen al cambiar los valores de estas variables; el paradigma ha sido fundamental en las obras de psicólogos que tienen tan poco en común como, por ejemplo, Skinner y Vygotsky[6].

---

[5] Este análisis del aprendizaje de conceptos está, por lo general, de acuerdo con autores como Bruner, Goodnow y Austin (1956), como ocurre con el énfasis puesto en el carácter inferencial de las computaciones que constituyen la base de los resultados positivos en las situaciones de aprendizaje de conceptos.

[6] Aunque quizá a Skinner no le gustaría verlo formulado de esta manera. Parte del análisis conductista radical del aprendizaje es el intento de reducir el aprendizaje de conceptos al «aprendizaje por discriminación»; es decir, insistir en que *lo que* aprende el organismo en la situación de aprendizaje de conceptos es a producir *la respuesta designada*. Sin embargo, parece claro que la reducción debería proceder exactamente al revés: el paradigma de aprendizaje de conceptos y el paradigma de aprendizaje por discriminación *son* los mismos, y en ninguno de los dos la existencia de una respuesta indicada pasa de ser una conveniencia del experimentador; lo único que hace es brindar un procedimiento reglamentado en virtud del cual *S* pueda indicar qué clasificación le parece la correcta en una determinada fase dentro del proceso de aprendizaje.

Esto no es una afirmación metodológica sino empírica. Por varias razones es claro que el aprendizaje

Lo que quiero destacar en este momento es que sólo se ha llegado a proponer un tipo de teoría en relación con el aprendizaje de conceptos —en realidad, se podría decir que sólo sería concebible una única clase de teoría— y que esta teoría carece de coherencia a no ser que se dé un lenguaje del pensamiento. En este sentido, el análisis del aprendizaje de conceptos es como el análisis de la elección deliberada; no podemos comenzar a ver el sentido de los fenómenos a no ser que estemos dispuestos a considerarlos como computacionales y no podemos comenzar a ver el sentido de la opinión que los considera como computacionales a no ser que estemos dispuestos a suponer un sistema representacional de considerable fuerza en que se lleven a cabo las computaciones.

Conviene tener en cuenta, en primer lugar, que en cualquier ensayo *t* y en relación con una determinada propiedad *P,* la experiencia del organismo en el paradigma de aprendizaje de conceptos se representa adecuadamente como una matriz de datos en que las filas representan ensayos y las columnas representan la ejecución de la respuesta indicada, la presencia o ausencia de *P,* y el carácter de la señal de error[7]. Así:

ENSAYO	EJECUCION DE LA RESPUESTA INDICADA	PROPIEDAD *P* PRESENTE	VALOR DE LA SEÑAL DE ERROR
1	sí	sí	menos
2	no	no	menos
3	sí	no	más

de conceptos (en el sentido de aprender qué categorización de los estímulos es la correcta) puede realizarse, y se realiza normalmente, en ausencia de respuestas indicadas específicas —de hecho, en ausencia de todo tipo de respuesta—. Los aficionados a la naturaleza creo que llegan a distinguir un roble de un pino, y muchos de ellos lo hacen probablemente sin que se les enseñen explícitamente los criterios que permiten establecer la distinción. Esto es verdadero aprendizaje de conceptos, sin que se dé una respuesta distintiva, ni siquiera en los amantes de la naturaleza, que se suela hacer cuando y sólo cuando se ve un roble.

En este sentido tenemos muchas pruebas experimentales. Tolman (1932) demostró que lo que aprende una rata cuando aprende qué giro se recompensa en un laberinto en forma de T *no* es específico del sistema de respuesta que utiliza para dar el giro. Brewer (todavía sin publicar), en un estudio reciente de las obras que estudian el condicionamiento en los seres humanos, da argumentos convincentes de que la respuesta designada se puede desligar de los estímulos de criterio sencillamente dando instrucciones al sujeto para que la desligue («Por favor, a partir de ahora *no* clasifique los círculos rojos con los cuadrados negros»). En resumen, no se da el caso de que el aprendizaje consista en establecer conexiones entre clases específicas de estímulos y clases específicas de respuestas. Lo que ocurre es a) que *S* puede utilizar muchas veces lo que ha aprendido para conseguir una correspondencia entre la presencia de estimulación-criterio y la producción de una respuesta designada; b) que muchas veces es experimentalmente conveniente exigirlo que lo haga, dándose así un procedimiento sencillo para que *E* determine cuáles son las propiedades de los estímulos que *S* cree que sirven de criterio, y c) que los *S*s continúan así con tal que se les motive adecuadamente. En esto, como en todo lo demás, lo que hace el sujeto está determinado por sus creencias y por sus preferencias.

[7] En teoría podría necesitarse una matriz de tres valores pues, en un ensayo determinado cualquiera, es posible que el organismo no haya observado, o que haya observado y olvidado, si se realizó la respuesta indicada, si estaba presente *P,* o cuál era el valor de la señal de error. Este tipo de sutilezas las dejaré de lado por regla general. Sólo hago mención de ello para insistir en que es la representación interna de sus experiencias (y no los hechos objetivos al respecto) hecha por el organismo lo que está inmediatamente implicado en la causalidad de la conducta.

Expuesto de esta manera, parece claro que el problema con que se enfrenta el organismo en el ensayo *t* es el de elegir un valor de *P* para el que, en el caso ideal, la última columna de la matriz sea positiva cuando y sólo cuando lo sean las dos primeras columnas, y que sea tal que la matriz siga manteniendo esa correspondencia para cualquier valor (razonable) de $t_n > t$. Este es el sentido en que lo que se aprende en el aprendizaje de conceptos «va más allá» de lo que se da en los datos experimentales. Lo que tiene que hacer el organismo para actuar con acierto es extrapolar una generalización (todos los estímulos positivos son estímulos *P*) basándose en algunos casos que se conforman a la generalización (los primeros *n* estímulos positivos fueron estímulos *P*). Se trata, en resumidas cuentas, de una extrapolación inductiva, y la extrapolación inductiva presupone: a) una fuente de hipótesis inductivas (en el caso presente, una serie de candidatos a *P*), y b) un criterio métrico de confirmación tal que la probabilidad de que el organismo acepte (por ejemplo, actúe sobre) un valor determinado de *P* en *t* sea una función razonable de la distribución de entradas en la matriz de datos correspondiente a los ensayos anteriores a *t*.

Existen, naturalmente, muchísimos procedimientos para explicitar los detalles de este tipo de modelo. Por ejemplo, hay muchas razones para creer que los distintos valores de *P* suelen comprobarse por lo general en un orden determinado; e incluso, que la elección de *P* puede estar determinada muy sutilmente por el carácter de los valores de *P* previamente evaluados y rechazados y por la configuración concreta de la matriz de datos correspondiente a esos valores. Pero, comoquiera que sean los detalles, lo que parece totalmente claro es que la conducta del organismo dependerá de la relación de confirmación entre los datos y las hipótesis, de forma que las explicaciones de su conducta requieran información sobre la manera en que, en el curso del aprendizaje, se representan los datos y las hipótesis.

¿Por qué resulta esto así de claro? Fundamentalmente, porque una de las características diferenciadoras del aprendizaje de conceptos es la *no-arbitrariedad* de la relación entre lo que se aprende y el carácter de las experiencias que ocasionan el aprendizaje. (Compárese esto con el caso de adquirir el latín tomando pastillas). Es decir, lo que una teoría del aprendizaje de conceptos tiene que explicar es por qué son las experiencias de *x* que son *F* (y no, por ejemplo, las experiencias de *x* que son *G*) las que llevan al organismo, en último término, a la creencia de que todas las *x* son *F*. *Podemos* explicar esto si suponemos: a) que el organismo *representa* las experiencias relevantes como experiencias de *x* que son *F;* b) que una de las hipótesis mantenidas por el organismo en relación con su entorno es la hipótesis de que quizá todas las *x* sean *F*, y c) que el organismo utilice en la fijación de sus creencias, una regla de confirmación que diga (en términos *muy* aproximativos) que el que todas las *x* observadas sean *F* es, *ceteris paribus,* un motivo para creer que todas las *x* son *F*. Para decirlo en términos más suaves, parece improbable que cualquier teoría radicalmente incompatible con los puntos (a-c) pueda explicar la no-arbitrariedad de la relación existente entre lo que se aprende y las experiencias que ocasionan el aprendizaje[8].

---

[8] He insistido deliberadamente en las analogías existentes entre la teoría de la confirmación inductiva y la teoría de la fijación de la creencia. Pero *no* es mi intención apoyar la opinión (que podrían insinuar ejemplos como el punto c)) de que la confirmación de las hipótesis universales en la ciencia es normalmente un proceso de simple generalización a partir de casos sueltos. En este sentido, tampoco trato de respal-

En resumen, el aprendizaje de conceptos exige un análisis que lleve a determinar una relación de confirmación entre las contingencias de recompensa observadas y extrapoladas, y esto es ya admitir un sistema representacional en que se manifiestan las observaciones y las extrapolaciones aspirantes y en que se computa el grado de confirmación. Sin embargo, existe una forma más sutil en que la extrapolación inductiva presupone un sistema representacional, y se trata de un tema que merece nuestra atención.

La extrapolación inductiva es una forma de inferencia no demostrativa. En relación con el tema que nos ocupa, esto quiere decir que, en un determinado ensayo *t*, habrá un número indefinido de valores no equivalentes de *P* que sean «compatibles» con la matriz de datos hasta *t*. Es decir, habrá un número indefinido de valores de *P* de forma que, en todos los ensayos anteriores a *t*, se recompense la respuesta indicada si y sólo si la propiedad *P* se pone de manifiesto en el estímulo, pero donde cada valor de *P* «predice» un diferente emparejamiento de respuestas y recompensas en los ensayos siguientes. Evidentemente, para que el organismo extrapole de sus experiencias, necesitará alguna forma de elegir entre estos valores de *P* de número indefinido. También es evidente que no se puede hacer la elección partiendo de la base de los datos disponibles hasta *t*, pues la elección que se debe hacer es precisamente entre las hipótesis, todas las cuales predicen los *mismos* datos hasta *t*.

Es ésta una situación muy común en las discusiones sobre la inferencia inductiva dentro de la filosofía de la ciencia. El argumento clásico es obra de Goodman (1965), quien señalaba que, para una serie fija de observaciones de esmeraldas verdes, resultarán compatibles con los datos tanto la hipótesis de que todas las esmeraldas son verdes como la hipótesis de que todas las esmeraldas son «*verzules*» *. (Una forma de definir el predicado «*verzul*» es la siguiente: Una esmeralda es verde si y sólo si [está en la muestra de datos y es verde] o [no está en la muestra de datos y es azul]). Sin embargo, una de las afirmaciones de Goodman es que existe un número indefinido de maneras de construir predicados que tengan en común las propiedades contrainductivas que presenta «*verzul*». Como ambas hipótesis son compatibles con los datos, el principio que permita discriminar entre ellas debe apelar a algo distinto de las observaciones de esmeraldas verdes.

---

dar la opinión, materializada en el programa de la «lógica inductiva», de que la confirmación se puede reconstruir generalmente en cuanto relación «formal» entre hipótesis y datos. Por el contrario, parece que el nivel de confirmación de una hipótesis científica suele ser sensible a una variedad de consideraciones *in*formales en relación con la economía, plausibilidad, fuerza de convicción y productividad generales de la teoría en que se enmarca la hipótesis, dejando de lado la existencia de teorías opuestas.

Puede ocurrir también que la fijación de la creencia sea sensible a este tipo de consideraciones «globales». Pero aun en ese caso las perspectivas de una teoría formal de la creencia me parecen mucho mejores que las de una lógica inductiva. Para formalizar la relación de confirmación inductiva, deberíamos suministrar una teoría que escoja *la mejor* de las hipótesis (la hipótesis que se *debería* creer), dadas las pruebas con que se cuenta. Por el contrario, para formalizar la fijación de la creencia, sólo debemos desarrollar una teoría que, dadas las pruebas existentes, elija la hipótesis en que *cree* el organismo. En la medida en que esto resulte imposible, es imposible considerar el aprendizaje como un proceso computacional; y, para bien o para mal, la suposición en que se basa este libro es que las explicaciones computacionales de los organismos no van a venirse abajo.

* Utilizamos esta palabra porque responde, en forma y contenido, a la palabra original del autor: «grue». Se trata de una fusión de elementos procedentes de las palabras gr(een) [= verde] y (bl)ue [= azul]. (N. del T.).

La forma de solucionar este rompecabezas es suponer que las extrapolaciones de los datos reciben una ordenación a priori según un *criterio de simplicidad,* y que ese criterio prefiere «todas las x son verdes» a «todas las x son *"verzules"*» como extrapolación de un conjunto de datos compatible con ambas[9]. En el caso presente esto significa que la decisión de que un determinado valor de P se confirma en relación a una determinada matriz de datos debe estar determinada no únicamente por la distribución de las entradas en la matriz, sino también por la sencillez relativa de P. Esta conclusión parece irresistible, dado el carácter no demostrativo de las extrapolaciones implicadas en el aprendizaje de conceptos. Sin embargo, tiene consecuencias inmediatas para la afirmación general de que las teorías del aprendizaje de conceptos son incoherentes a no ser que presuponga que el organismo puede contar con un sistema representacional.

Lo importante es que el criterio de simplicidad debe ser sensible a la *forma* de las hipótesis a las que se aplica, es decir a su sintaxis y vocabulario[10]. Por lo que podemos saber, es posible conseguir una ordenación a priori de las hipótesis solamente si tenemos en cuenta la forma en que se expresan las hipótesis. Necesitamos una ordenación de esas características si queremos presentar una explicación coherente del orden en que se seleccionan los valores de P en la situación de aprendizaje de conceptos. Pero esto quiere decir que una teoría del aprendizaje de conceptos tendrá que ser sensible a la forma en que el organismo representa sus hipótesis. Pero la idea de que el organismo representa sus hipótesis de una forma u otra (por ejemplo, en un vocabulario u otro, en una sintaxis u otra) *es* precisamente la idea de que el organismo posee un sistema representacional.

En realidad, este argumento hace una exposición del caso que resulta demasiado débil. En la formalización de la inferencia científica un criterio de simplicidad distingue entre hipótesis que son compatibles con los datos pero que dan lugar a predicciones diferentes en relación con los casos *no* observados. Nuestra idea central es que las observaciones correspondientes son probablemente válidas en el caso especial en que las hipótesis sean valores de P y los datos sean los valores observados de la señal de error. Existe, sin embargo, un aspecto en que el caso de la inferencia científica difiere de las extrapolaciones que intervienen en el aprendizaje de conceptos. Parece que no hace falta el criterio de simplicidad utilizado en la evaluación de las teorías científicas para distinguir entre hipótesis *equivalentes.* Dicho a la inversa, dos hipóte-

---

[9] Considero que esto es muy común entre los filósofos de la ciencia. En lo que no están de acuerdo es en la forma de describir la diferencia entre predicados como *verzul* (que no son del agrado del criterio de simplicidad) y predicados como verde (que sí lo son); y tampoco en la forma de justificar la adopción de un criterio de simplicidad que discrimine de esa manera.

[10] Ideas como la de atrincheramiento, por ejemplo, se definen según los *predicados* de una ciencia. Si «verde» está más atrincherado que «verzul», se debe probablemente a que hay leyes expresadas desde el punto de vista del primero y en cambio no hay leyes expresadas desde el punto de vista del segundo. (Para exposición más detallada véase Goodman, 1965). Como es lógico, podría evitarse esta conclusión definiendo la simplicidad, atrincheramiento y nociones relacionadas en relación con las *propiedades* (y no con los predicados). Pero aun cuando fuera *posible* hacerlo, parecería ser un paso por un camino equivocado: en la medida en que se quiere que los procesos psicológicos se manifiesten como procesos *computacionales,* se quiere que las reglas de la computación se apliquen formalmente a los objetos que están en sus dominios. Mi objetivo en este libro, vuelvo a decirlo, no es *demostrar* que los procesos psicológicos son computacionales, sino considerar las consecuencias de suponer que lo sean.

sis son idénticas, en relación con la formalización de las inferencias científicas, si se predicen las mismas extrapolaciones de la matriz de datos y son igualmente complejas. De los pares de hipótesis que son idénticos en este sentido pero difieren en cuanto a su formulación, se dice que son «variantes notacionales» de la misma teoría.

Sin embargo, existen numerosas pruebas de que la ordenación a priori de los valores de *P* a los que se recurre en el aprendizaje de conceptos distingue entre hipótesis que son, en este sentido, variantes notacionales mutuas; es decir, la ordenación de valores de *P* impone exigencias más *fuertes* sobre la forma de una hipótesis que la métrica de la sencillez.

Por ejemplo, es un hecho comprobado que los *S*s prefieren representaciones conjuntivas afirmativas de la matriz de datos a las representaciones negativas o disyuntivas. (Véase Bruner et al., 1956). Así, los sujetos de una tarea de aprendizaje de conceptos tendrán por lo general menos dificultades para aprender a clasificar todos los triángulos rojos juntos que para aprender a clasificar juntas todas las cosas que *no sean* triángulos o todas las cosas que sean o triángulos o rojas. Sin embargo, las hipótesis conjuntivas afirmativas son interdefinibles con las hipótesis disyuntivas negativas; el sujeto que está eligiendo todos y sólo los triángulos rojos como ejemplos de estímulos positivos está eligiendo, ipso facto, todas y únicamente las cosas que son (no triángulos o no rojas) como ejemplos de los estímulos negativos[11]. Lo que establece la diferencia en la actuación del sujeto es cuál de estas opciones considera él que está haciendo; es decir, la forma en que representa sus opciones. Los *S*s que hablan de una hipótesis conjuntiva afirmativa aprenden por lo general más rápidamente que los que no lo hacen[12]. Esto resulta perfectamente inteligible si partimos de la suposición de que la misma hipótesis puede recibir diferentes representaciones internas y que las preferencias a priori del sujeto son sensibles a estas diferencias. Lo cual parece que no se puede entender con ninguna otra explicación.

Hemos estado considerando algunas de las formas en que la idea de que la tarea de aprendizaje de conceptos conlleva esencialmente una extrapolación inductiva nos

---

[11] Lo importante es, naturalmente, que la «elección» en el primer caso es opaca y en el segundo es transparente. Quizá no resulte sorprendente que lo que se elige opacamente se elija según una representación.

[12] Por ejemplo, Wason y Johnson-Laird (1972) describen un experimento en que se presentaba a los *S*s matrices de datos y se les pedía que articularan las extrapolaciones adecuadas. La predicción básica, que se vio confirmada, era que «sería más fácil formular los conceptos que eran esencialmente conjuntivos por la forma que los conceptos que eran esencialmente disyuntivos en cuanto a la forma, y que siempre que se negaba un componente se producía un ligero aumento de la dificultad» (p. 70). Observan que el orden de dificultad que obtuvieron preguntando al sujeto que afirmara la generalización pertinente «se conforma al orden obtenido cuando los sujetos tienen que *aprender* conceptos en la forma convencional» (p. 72), es decir, en la tarea de aprendizaje de conceptos. Lo importante a tener en cuenta es que, dado que la conjunción es interdefinible con la negación y la disyunción, no hay ningún concepto que sea, *estrictamente hablando,* esencialmente conjuntivo o esencialmente disyuntivo. Estrictamente hablando, los conceptos no *tienen* formas, aunque las representaciones de los conceptos las tengan. Lo que Wason y Johnson-Laird quieren decir al hablar de concepto conjuntivo es, como ellos mismos señalan con especial cuidado, que se puede expresar mediante una fórmula (relativamente) económica *en el sistema representacional que está utilizando el sujeto* (en el caso actual, en inglés). Lo que demuestra realmente el experimento es que la utilización de esta representación facilita la actuación del sujeto; por eso, las formulaciones de una hipótesis que son, en el sentido descrito anteriormente, meras variantes notacionales mutuas, pueden sin embargo ser asequibles de distinta manera en cuanto extrapolaciones de una matriz de datos.

obliga a postular un sistema representacional en que se lleven a cabo las inducciones pertinentes. Creo conveniente insistir en que no se ha propuesto jamás otra opinión alternativa sobre el aprendizaje de conceptos, aunque existen vocabularios alternativos para formular el punto de vista que acabamos de exponer. Por ejemplo, muchos psicólogos utilizan la idea de fuerza del hábito (o fuerza de asociación) donde yo he utilizado la noción de grado de confirmación de una hipótesis. Pero una vez que se ha reconocido que todo constructo de esta naturaleza debe definirse en relación con las posibles extrapolaciones de la matriz de datos (y no en relación con pares E-R; véase nota 6), el problema que queda es totalmente terminológico. Una teoría que determina cómo varía la fuerza del hábito en función del reforzamiento (o que determina la fuerza de asociación en función de la frecuencia de asociación, etc.) *es* precisamente una lógica inducíva, en que la función de confirmación se articula por cualquiera de las leyes de reforzamiento/asociación que se presupongan.

De la misma manera, algunos psicólogos preferirían hablar de una teoría que determina el orden en que se comprueban los valores de *P*. Pero también en este caso el problema es meramente terminológico. Una teoría que determine aquello a lo que el organismo atiende en el ensayo *t*, por el mismo hecho prevé el parámetro del estímulo que se extrapola en *t*. Por consiguiente tiene que ser sensible a las propiedades de la matriz de datos, y de las hipótesis previamente contempladas, que afecten al orden en que se comprueban los valores de *P*, y al ordenamiento a priori de los valores de *P* que determina su relativa complejidad. Tanto si *llamamos* a esto teoría de la atención como si no, la función del constructo es precisamente prever cuáles son las extrapolaciones de la matriz de datos que ensayará el organismo y el orden en que lo hará.

Finalmente, hay psicólogos que prefieren describir el organismo como si «hiciera un muestreo» de las propiedades del estímulo en vez de construir hipótesis sobre cuáles de estas propiedades tienen valor de criterio para la clasificación. Pero la noción de propiedad es propia de la primera clase de teoría. En el sentido no propio de «propiedad», todo estímulo tiene una infinidad de propiedades de las cuales hay un subconjunto infinito que nunca se toma como muestra. Por otra parte, las propiedades que sí se toman como muestra son necesariamente una selección de aquellas que el organismo es capaz de representar internamente. Por eso, el hablar sobre hacer un muestreo de las propiedades del estímulo y el hablar sobre proyectar hipótesis sobre estas propiedades son dos formas de afirmar lo mismo.

Resumiendo: Por lo que sabemos, el aprendizaje de conceptos es esencialmente una extrapolación inductiva, por lo que una teoría del aprendizaje de conceptos tendrá que presentar los rasgos característicos de las teorías inductivas. En concreto, el aprendizaje de conceptos presupone un formato para representar los datos experienciales, una fuente de hipótesis para predecir los datos futuros, y una métrica que determine el nivel de confirmación que un cuerpo determinado de datos confiere a una hipótesis determinada. Nadie, que yo sepa, lo ha puesto en duda, aunque supongo que muchos psicólogos no han llegado a darse cuenta de qué era lo que no ponían en duda. Pero aceptar que el aprendizaje que «va más allá de los datos» implica una inferencia inductiva significa que uno acepta un lenguaje en que se llevan a cabo las inducciones, pues: a) un argumento inductivo sólo se justifica en la medida en que las afirmaciones de observación que constituyen sus premisas confirman las hipótesis

que constituye la conclusión; b) el que esta relación de confirmación se dé entre las premisas y la conclusión depende, al menos en parte, de la *forma* de las premisas y de la conclusión, y c) la idea de «forma» sólo se define en relación con los objetos «lingüísticos», es decir, con las representaciones.

Para terminar este capítulo quiero señalar que son estas mismas las conclusiones a las que se llega cuando se comienza a pensar sobre la estructura de las teorías de la percepción.

En primer lugar, existe una analogía evidente entre las teorías del aprendizaje de conceptos del tipo que acabo de exponer y las teorías clásicas de la percepción dentro de la línea empirista. Según estas últimas, la percepción es esencialmente cuestión de resolución de problemas, en que la forma del problema es prever el carácter de la experiencia sensorial futura, dado el carácter de las sensaciones pasadas y actuales en cuanto datos. Concebidos de esta manera, los modelos de la percepción tienen la misma estructura general que los modelos del aprendizaje de conceptos: Hace falta una forma canónica para la representación de los datos, hace falta una fuente de hipótesis para la extrapolación de los datos, y hace falta una métrica de confirmación para elegir entre las hipótesis.

Como algunos de los empiristas consideraron que su proyecto era la formalización de los *argumentos* perceptivos —es decir, de aquellos argumentos cuya fuerza justifica nuestras pretensiones de conocimiento sobre los objetos de la percepción— elaboraron doctrinas bastante explícitas sobre las clases de representaciones que intervienen en las inferencias perceptuales. Es posible (y está dentro del espíritu de gran parte de la tradición empirista) considerar que estas doctrinas implican las teorías de los procesos computacionales que subyacen a la integración perceptual. Sin embargo, es bien conocido que en muchos aspectos las explicaciones empiristas de las inferencias perceptuales dan lugar a una psicología de dudoso valor, cuando se realiza de esa manera. Por ejemplo, en algunas ocasiones se ha supuesto que las premisas de las inferencias perceptuales estaban representadas en un lenguaje «de datos sensoriales» cuyas fórmulas se pensaba que tenían ciertas propiedades muy peculiares: por ejemplo, que las afirmaciones del dato sensorial son de alguna manera incorregibles, que todas las afirmaciones empíricas sólo se pueden descomponer de una manera en afirmaciones del dato sensorial; que cada afirmación del dato sensorial es lógicamente independiente de las demás, y así sucesivamente.

Para muchos de los empiristas, el rasgo determinante de este lenguaje de datos era que sus expresiones referentes sólo podían referirse a los «qualia». Si las afirmaciones de datos sensoriales eran curiosas, lo eran porque los qualia todavía lo eran más. A la inversa, el lenguaje en que se formulan las hipótesis perceptuales se identificaba con el «lenguaje del objeto físico», haciendo con ello la distinción entre lo que se siente y lo que se percibe coextensiva con la distinción entre qualia y cosas. Las redescripciones de los campos sensoriales desde el punto de vista del objeto físico podrían servir para la predicción de sensaciones futuras porque, según esta opinión, aceptar una descripción de la propia experiencia en un lenguaje de objeto físico significa lógicamente aceptar afirmaciones (al menos hipotéticas) sobre las experiencias todavía por venir. En términos aproximados se podría decir que las afirmaciones del dato sensorial constituyen un apoyo inductivo para las afirmaciones del objeto físico, y las afirmaciones del objeto físico implican afirmaciones sobre futuras sensaciones.

De esta manera, al pasar por inferencia de las sensaciones a las percepciones, se acepta un «riesgo inductivo», y el problema que se plantea al perceptor es el de comportarse racionalmente frente a este riesgo. Es decir, dada una descripción de la experiencia formulada en el lenguaje de la sensación, el sujeto debe elegir de alguna manera la *re*-descripción en términos del objeto físico que se vea más confirmada por las experiencias. Sólo de esta manera puede estar razonablemente seguro de que la mayoría de las expectativas sobre experiencias futuras o hipotéticas a que le impulsan sus juicios perceptuales tienen probabilidades de ser verdaderas.

Si describo mi experiencia actual en términos de manchas de color, texturas, olores, sonidos, etc., no me veo obligado a aceptar predicciones sobre el carácter de mis experiencias anteriores o futuras. Pero si la describo en términos de mesas y sillas y sus familias lógicas, en ese caso me veo comprometido a aceptar tales predicciones, pues no hay nada que pueda ser una mesa o una silla a no ser que se comporte a lo largo del tiempo según lo que podríamos considerar una manera razonable de ser mesa o silla. Así, si afirmo que lo que veo es una mesa, me estoy comprometiendo (implícitamente) con su comportamiento pasado y futuro; en concreto, estoy dando garantías sobre las sensaciones que proporcionará o proporcionaría. Ese es el asunto.

Es sabido por todos que esta explicación de la percepción ha sufrido una grave derrota por obra de los epistemólogos y de los psicólogos de la Gestalt. Hoy día resulta difícil imaginarse qué pasaría si las fórmulas de un sistema representacional se vieran con los privilegios que se atribuyeron a las fórmulas en el lenguaje del dato sensorial. Tampoco es fácil imaginar una forma de describir los qualia que obligara a admitir que toda la información perceptiva está mediada por la sensación de los mencionados qualia. Y tampoco parece que tenga mucho sentido negar que lo que vemos son generalmente *cosas;* no, al menos, si la alternativa es que lo que vemos son generalmente manchas de colores y sus bordes.

Esta línea de crítica es demasiado conocida como para que tengamos que repetirla aquí. Creo que resulta perfectamente convincente. Pero, sin embargo, opino que el núcleo de la teoría empirista de la percepción es inevitable. En concreto, me parece que las siguientes afirmaciones sobre la psicología de la percepción son casi totalmente verdaderas y dentro del espíritu de la teoría empirista:

1. La percepción implica por norma general la formación y confirmación de hipótesis.
2. Los datos sensoriales que confirman una determinada hipótesis perceptual se suelen representar internamente en un vocabulario que se puede considerar empobrecido si se lo compara con el vocabulario en que se formulan las mismas hipótesis.

Antes de explicar por qué considero ciertos estos aspectos del tratamiento empirista de la percepción quiero referirme brevemente a lo que pienso fue el error de los empiristas.

Estoy interpretando la teoría empirista de la percepción como si cumpliera una doble misión: ser una explicación de la justificación de las creencias perceptivas y ser una psicología de la integración de los preceptos. Creo que muchos de los empiristas iban por este camino. Pero también es claro que cuando surgió un conflicto entre lo

que exigía la psicología y lo que parecía exigir la epistemología, fueron estas últimas exigencias las que configuraron la teoría.

Por ejemplo, la afirmación de la incorregibilidad de las afirmaciones del dato sensorial no respondía a ninguna idea psicológica concreta, sino más bien a la supuesta necesidad de aislar el riesgo inductivo en un nivel epistémico distinto de aquel en que se especificaban los datos. La idea era, poco más o menos, que no podíamos saber si las afirmaciones del objeto físico eran verdaderas a no ser que estuviéramos seguros de los datos correspondientes a dichas afirmaciones, y no podríamos estar *seguros* de las afirmaciones de los datos si fuera posible que algunos de ellos fueran falsos. La certeza, por así decirlo, se transmite en dirección ascendente, desde los datos a los juicios perceptivos que se apoyan en ellos. De la misma manera, las experiencias de los qualia tienen que ser hechos conscientes porque las afirmaciones que confirman estas experiencias son las premisas de los argumentos cuyas conclusiones son las afirmaciones del objeto físico que creemos explícitamente. Si estos argumentos van a ser nuestra justificación para creer tales afirmaciones, sería conveniente tener a mano sus premisas para poder citarlas.

Es muy probable que todo esto resulte muy confuso. La justificación es una noción mucho más pragmática de lo que sugiere el análisis empirista. En concreto, no hay ninguna razón para que la dirección de todos los argumentos justificatorios sea ascendente y parta de premisas epistemológicamente inatacables. ¿Por qué no puede una de mis afirmaciones del objeto físico estar justificada por recurso a otra, y ésta por recurso a una tercera, y así sucesivamente? Lo que exige el argumento justificatorio no es que algunas de las creencias sean incuestionables sino, como máximo, que algunas de ellas no sean cuestionadas (de facto). Lo que *no se puede* hacer es justificar todas mis creencias *a la vez*. Y lo que no se puede, no se puede.

Pero aunque piense que la idea de *la* dirección de la justificación está en gran parte desorientada, la idea de que hay una dirección del flujo de la información *en la percepción* está casi con toda seguridad bien orientada, a pesar de que los argumentos son empíricos más que conceptuales.

En primer lugar, parece claro que las interacciones causales entre el organismo y su entorno deben contribuir a la etiología de todo lo que alguien se sienta inclinado a llamar conocimiento *perceptivo*. En la medida en que esto es cierto, existe gran cantidad de información empírica sobre el carácter de estas interacciones.

Según es sabido, toda información que el organismo recibe sobre su entorno como consecuencia de estas interacciones debe estar mediada por la actividad de un *mecanismo sensorial* u otro. Por mecanismo sensorial entiendo aquel que responde a las *propiedades físicas* de los hechos del entorno. Por propiedad física entiendo la designada por un término de clase natural en una ciencia física (supuestamente terminada) (sobre la noción de término de clase natural, véase la segunda parte de la introducción). Hace falta una explicación relativamente amplia para ver lo que significa *mediada por,* pero como primera aproximación lo que quiero decir es que la operación de un mecanismo sensorial al responder a una propiedad física de un hecho del entorno es condición empíricamente necesaria para la percepción por el organismo de *cualquier* propiedad de ese hecho del entorno.

Supongamos, por ejemplo, un mecanismo sensorial que está representado por una función característica, de manera que el valor de la función sea 1 en todos los

casos en que se excita el mecanismo y 0 en los demás. Entonces, como es bien sabido, podemos elaborar una teoría que prevea los valores de dicha función a lo largo del tiempo solamente si tenemos en cuenta las propiedades físicas de los inputs que recibe el mecanismo. Y podremos predecir el análisis perceptual que el organismo atribuirá a un hecho ambiental determinado únicamente si sabemos a qué propiedades físicas de tal hecho han respondido los mecanismos sensoriales del organismo. (Así, por ejemplo, para predecir el estado de excitación del sistema auditivo humano, necesitamos información sobre el análisis espectral de las formas ondulatorias que afectan al organismo. Y para predecir el tipo de frase a que se atribuirá perceptualmente una elocución dada, debemos saber al menos cuáles son las propiedades auditivas de la elocución que se han detectado).

Quiero insistir en que esto es un hecho *empírico,* aunque no un hecho *sorprendente.* Podemos imaginar un organismo (por ejemplo un ángel o un vidente) cuyo conocimiento perceptivo *no* esté mediado por el funcionamiento de los mecanismos sensoriales: lo único que ocurre es que, por lo que nosotros sabemos, no existen tales organismos, o, si los hay, los psicólogos todavía no los han encontrado. En todos los casos conocidos, la percepción depende del funcionamiento de mecanismos cuyos estados de excitación se pueden prever a partir de las descripciones físicas de su input y de ninguna otra forma.

Desde el punto de vista del flujo de la información esto quiere decir que hay un mecanismo sensorial que actúa para asociar una excitación física dada (como input) con una descripción física dada (como output); es decir, un mecanismo sensorial es un dispositivo que dice «sí» cuando es excitado por estímulos que presentan ciertos valores determinados de los parámetros físicos y «no» en los demás casos[13]. En concreto, dicho mecanismo no se preocupa de ninguna propiedad que los hechos del entorno *no* lleguen a compartir con tal que los hechos tengan en común las propiedades físicas relevantes, y tampoco se ocupa de las propiedades no físicas que tengan en común los hechos del entorno con tal que no lleguen a compartir las propiedades físicas relevantes. En este sentido, la excitación de un mecanismo sensorial codifica la presencia de una propiedad física. (Si el sistema auditivo es un mecanismo cuyos estados de excitación son específicos respecto a los valores de frecuencia, amplitud, etc., de los hechos ambientales que le influyen causalmente, podríamos también considerar que el output del sistema es una descripción codificada del entorno atendiendo a esos valores. En realidad, sería mejor considerarlo de esta manera si se intenta representar la integración de los perceptos auditivos como un proceso *computacional.*) Pero si esto es cierto, y si es también cierto que toda información perceptual del organismo sobre su entorno está mediada por la actuación de sus mecanismos sensoriales, se deduce que los análisis perceptuales deben ser sensibles de alguna manera a la informa-

[13] Por exigencias de la exposición, estoy prescindiendo de la (seria) posibilidad empírica de que algunos o todos los mecanismos sensoriales tengan valores de output entre 0 y 1. Aquí entran en juego los problemas sobre la «digitalidad» de las diversas etapas del procesamiento cognitivo; pero, aunque se trata de problemas interesantes e importantes, no afectan a las cuestiones más generales. Baste decir que la cuestión no es si los outputs de los mecanismos sensoriales son continuos bajo una descripción física, sino más bien si los valores intermedios de excitación comunican información que se utiliza en etapas posteriores del procesamiento. No sé cuál será la respuesta a esta cuestión y no trato de excluir la posibilidad de que la respuesta sea diferente para las diferentes modalidades sensoriales.

cion sobre los valores de los parámetros físicos de los hechos del entorno que suministran los mecanismos sensoriales[14].

Ese *es,* supongo, el problema de la percepción en la medida en que el problema de la percepción constituye un problema en psicología. Pues aunque la información suministrada por las interacciones causales entre el entorno y el organismo es información sobre las propiedades físicas en *primer* lugar, en *último* término puede (desde luego) ser información sobre cualquier propiedad que el organismo perciba en el entorno. En una primera aproximación, los outputs de los mecanismos sensoriales son considerados merecidamente como descripciones físicas, pero no hace falta que los juicios perceptuales estén articulados en el vocabulario de estas descripciones. Por lo general no lo *están:* Un juicio perceptivo paradigmático es: «Hay un petirrojo en el césped» o «Por el reloj, veo que es la hora de tomar el té».

Creo que la cuestión de si los procesos psicológicos son procesos computacionales es una cuestión empírica. Pero si es que lo son, entonces lo que debe ocurrir en la percepción es que una descripción del entorno que *no* esté formulada en un vocabulario cuyos términos designen valores de variables físicas se computa de alguna manera partiendo de la base de una descripción que *sí* está formulada en dicho vocabulario. Parece que esto es posible, ya que el análisis perceptual de un hecho está determinado no solamente por la información sensorial, sino también por el conocimiento previo que el

---

[14] Conviene destacar que la presente exposición de los sistemas sensoriales, como la mayor parte de las teorías psicológicas de este capítulo, está muy idealizada. Así, «desde el punto de vista físico los receptores sensoriales son transductores, es decir, convierten la forma concreta de energía a que se adapta cada uno en la energía eléctrica del impulso nervioso» (Loewenstein, 1960). Pero, naturalmente, eso no implica que los sensores sean transductores *perfectos,* es decir, que su output se pueda prever *simplemente* mediante una determinación de las energías físicas que inciden en ellos. Por el contrario, hay pruebas de que estas determinaciones pueden estar influenciadas por cualquiera de las siguientes variables, o por todas ellas.

i. Las células de los sistemas sensoriales tienen un ciclo característico de inhibición y de aumento de la sensibilidad consiguiente a cada disparo. Por eso, los efectos de los estímulos no son independientes de los efectos de las estimulaciones previas a no ser que el intervalo entre estímulos sea grande en comparación con el tiempo que suele ocupar dicho ciclo.

ii. Las células de la periferia sensorial pueden estar interconectadas de tal manera que la excitación de una de ellas inhibe el disparo de las demás. Esta inhibición «lateral» mutua de los elementos sensoriales se suele interpretar como si se tratara de un mecanismo de «agudizamiento»; quizá como parte de un sistema general de conversión de-análogo-a-digital (véase Ratliff, 1961).

iii. A cualquier distancia «de regreso» de la periferia del sistema sensorial existen probabilidades de encontrar elementos «lógicos» cuya activación se pueda considerar como codificación de funciones de Bool de la información del transductor primario (véase Letvin et al., 1961, Capranica, 1965).

iv. Puede haber una sintonía «centrípeta» central de las características de respuesta de los transductores periféricos, en cuyo caso el output de tales transductores puede variar según el estado motivacional, atencional, etc., del organismo.

v. Las células del sistema sensorial tienen actividad «espontánea», es decir, activación que *no* depende de los inputs estimuladores.

Por eso, un transductor sensorial puede divergir, en todos estos aspectos, de los mecanismos ideales contemplados en el texto; tampoco pretendo que esta enumeración sea completa. Pero, en cualquier caso, sigue siendo válida la afirmación central: en la medida en que el entorno contribuye a la etiología de la información sensorial, las uniformidades de su contribución parecen revelarse únicamente en su descripción física. O lo que sería equivalente a este respecto: en la medida en que la actividad de los mecanismos sensoriales codifica la información sobre el estado del entorno, lo que se codifica es el estado físico del entorno.

organismo aporta a la tarea. Los procesos computacionales de la percepción son principalmente los que intervienen en la integración de estas dos clases de información. Supongo que esto es lo que queda de la opinión empirista clásica de que la percepción implica una inferencia (no demostrativa) que parte de descripciones formuladas en un lenguaje relativamente pobre para llegar a conclusiones formuladas en otro lenguaje relativamente no empobrecido.

Con esto, no queda ya casi nada de la epistemología empirista. Por ejemplo, la descripción perceptualmente pertinente de la información sensorial no se da en el lenguaje, libre de teorías, de los qualia sino más bien en el lenguaje, cargado de teoría, de los valores de los parámetros físicos. (Esto es otra forma de deeir lo que ya he dicho más arriba: que, por lo que podemos saber, la única forma de presentar una explicación razonablemente sólida de la función característica de un mecanismo sensorial es considerar a sus inputs bajo una descripción física.) Por eso, no hay ninguna razón para creer que el organismo no se puede equivocar al aplicar una determinada descripción sensorial en un caso concreto. En este sentido, no hay ninguna razón para creer que los organismos son generalmente conscientes de los análisis sensoriales que imponen.

Esta distinción —entre la noción de mecanismo sensorial en cuanto fuente de un mosaico de experiencias conscientes a partir de las cuales se construyen los perceptos (por ejemplo, mediante procesos asociativos) y la noción de los sensores en cuanto transductores de la información del entorno que afecta a la integración perceptual— está ya consagrada en las obras de psicología. Insisten en ella incluso psicólogos como Gibson (1966), cuyo enfoque de la percepción no es, en conjunto, partidario del tipo de enfoque computacional de la psicología de que me estoy ocupando fundamentalmente. Para Gibson la percepción implica la detección de propiedades invariantes (generalmente relacionales) de los conjuntos de estímulos que actúan. Aparentemente presupone que todo percepto se puede identificar con dicha invariante si al menos la propiedad pertinente se describe con la suficiente abstracción[15]. Pero, aunque Gibson niega que los perceptos se construyan a partir de los datos sensoriales conscientes, parece afirmar de algún modo que la presencia de la invariante de estímulo pertinente debe deducirse a partir del output de información de los transductores sensoriales.

---

[15] No está clara la consideración que merece la afirmación de que hay invariantes del estímulo que corresponden a preceptos. Según una forma de interpretación, parecería constituir una verdad necesaria: dado que «percibir» es un verbo de consecución, tiene que haber al menos un rasgo invariante de todas las situaciones en que alguien perciba que una cosa es del tipo *t*; es decir, la presencia de una cosa del tipo *t*. Por otra parte, una afirmación *empírica* muy fuerte es que, en relación con cualquier tipo de cosa que se pueda percibir, existe un conjunto de propiedades *físicas* de tal manera que la detección de estas propiedades se pueda identificar plausiblemente con la percepción de una cosa de ese tipo. Esto último exige que la distinción entre cosas del tipo *t* y todas las demás cosas *sea una distinción física,* y, como vimos en la introducción, dicha conclusión *no* se sigue precisamente de la premisa de que los objetos del tipo *t* son objetos físicos.

La cuestión es si existen clases físicas que correspondan a clases perceptuales y eso, como venimos diciendo, es una cuestión empírica. La impresión que me producen las publicaciones que se ocupan del tema es que son más los casos en que no se da correspondencia que aquellos en que se da; que, en general, no se puede pensar que la percepción sea la categorización de las invariantes *físicas,* por muy abstractamente que se describan tales invariantes. (Para una exposición de la situación empírica en el campo de la percepción del habla, cf. Fodor et al., 1974.)

... Distinguiré entre el input al sistema nervioso que evoca la sensación consciente y el input que evoca la percepción. ... Un hecho indudable es que en algunas ocasiones es posible que se *detecte* algo sin el acompañamiento de las impresiones sensoriales. Un ejemplo es la detección visual de una cosa que está detrás de otra. ... Pero esto no significa que la percepción se pueda dar sin estimulación de los receptores; sólo quiere decir que los órganos de la percepción se ven en algunas ocasiones estimulados de tal manera que no se especifican en la conciencia. La percepción no puede darse... sin input; sólo puede ser así si con eso se quiere decir sin conciencia de la cualidad visual, auditiva, etc., del input. Un ejemplo de esto es el «sentido de los obstáculos» del ciego, que se siente como «visión facial» pero en realidad se trata de una detección auditiva del eco. El hombre ciego «siente» la pared que hay delante de él sin darse cuenta de cuál ha sido el sentido estimulado. En resumen, puede haber percepción sin sensación, pero no percepción sin información (p. 2).

Así, incluso para los psicólogos que piensan que las distinciones perceptuales son distinciones entre invariantes del estímulo (abstractas), es necesario resolver el problema de cómo se detectan estas invariantes; y parece que para resolverlo hace falta postular las mismas clases de inferencias basadas en los inputs que se suponen en las teorías empiristas. La diferencia consiste fundamentalmente en que los psicólogos actuales no suponen que las computaciones, o los datos sobre los que se definen, tengan que ser accesibles conscientemente[16].

Conviene insistir en que la afirmación de que los outputs de los mecanismos sensoriales no son, por lo general, accesibles conscientemente, se supone que es un resultado empírico y no una verdad de epistemología. Existen, por ejemplo, importantes pruebas empíricas de que una representación inicial de una señal del habla tiene que especificar sus relaciones formantes[17]. Sin embargo, los hablantes/oyentes no tienen acceso consciente a la estructura de formantes, y en este sentido, muy poco acceso consciente a cualquier otra propiedad acústica del habla. De hecho es muy probable

---

[16] Gibson se expresa algunas veces como si el problema de cómo se detectan las (supuestas) invariantes del estímulo se pudiera evitar distinguiendo entre el estímulo correspondiente a los *transductores sensoriales* (es decir, energías físicas) y el estímulo correspondiente a los *órganos perceptivos* (es decir, invariantes abstractas). Pero por ese camino podemos llegar a la trivialización. Si podemos utilizar la noción de estímulo para distinguir el input que llega a la retina (energía lumínica) del input que llega al sistema óptico (patrones de energía lumínica que presenta invarianzas relevantes, por ejemplo, para la explicación de las constancias perceptivas), ¿por qué no vamos a hablar también del estímulo relacionado con *todo el organismo* (es decir, de los perceptibles)? De esta manera, la respuesta a «¿Cómo percibimos las botellas?» sería: «Es necesario y suficiente para la percepción de una botella que se detecte la presencia del estímulo invariante *botella*». El problema de esta respuesta (que, por cierto, me suena muy dentro del estilo de Ryle) es, naturalmente, que el problema de cómo se detecta el estímulo invariante pertinente es el *mismo* problema de cómo se percibe una botella, por lo que no se ha conseguido ganar un palmo de terreno.

Todo lo cual demuestra, pienso yo, no que el problema psicológico de la percepción sea muy confuso, sino que el mero *enunciado* del problema exige la selección (y motivación) de un vocabulario propio para la representación de los inputs. He intentado demostrar que el vocabulario de los valores de los parámetros físicos *es* apropiado si partimos del supuesto plausible de que los transductores sensoriales detectan los valores de los parámetros físicos y de que todo el conocimiento perceptual está mediado por la actividad de los transductores sensoriales.

[17] Estoy dando por supuesto que las representaciones de un hecho del entorno que se asignan en el curso del análisis perceptual se computan serialmente. En realidad, bastará con una suposición no tan fuerte: es decir, que al menos *cierta* información sobre los parámetros físicos suele «entrar» antes de que se compute ninguna representación de nivel superior. Supongo que se trata de una afirmación que ningún psicólogo estaría dispuesto a negar.

que sea una verdad general el que, de las distintas redescripciones del input que subyacen a los análisis perceptuales, el grado de accesibilidad consciente de una representación se pueda predecir bastante bien a partir de su grado de abstracción respecto a lo que especifican los sensores. Esto es lo que tenían presente filósofos como Cassirer cuando señalaban que «oímos a través» de la elocución de una frase y llegamos a su significado; se nos da mucho mejor el informar del tipo sintáctico al que pertenece una elocución que el informar de las propiedades acústicas de la señal, y se nos da mucho mejor informar de los rasgos sintáticos que afectan al significado que de los que no lo afectan. Podríamos expresar esto diciendo que no oímos las relaciones formantes en las elocuciones de frases aun cuando oigamos las relaciones lingüísticas y aun cuando la estructura formante *(inter alia)* determine causalmente *cuáles* son las relaciones lingüísticas que oímos. Evidentemente, habría que calificar de lábil la determinación de las descripciones que son accesibles conscientemente. Los artistas y los especialistas en fonética aprenden conscientemente a advertir propiedades de su experiencia sensorial ante las que el profano es ciego y sordo. Este hecho no carece de interés, ni mucho menos; en el Capítulo 4 nos ocuparemos de algunas de sus consecuencias en relación con la teoría de la representación interna.

Hemos llegado, por tanto, a la idea de que la etiología de los análisis perceptuales implica una serie de redescripciones del entorno, y de que la descripción inicial en esta serie determina las propiedades físicas del entorno que son perceptivamente relevantes. La percepción debe incluir la formación y confirmación de hipótesis, ya que el organismo debe conseguir de alguna manera inferir la adecuada descripción del entorno que sea relevante para la tarea en cuestión partiendo de su descripción física y de toda la información anterior de que disponga sobre la estructura del entorno. Curiosamente, esta inferencia no es demostrativa: Por lo general, no existe ninguna conexión *conceptual* entre una categoría perceptual y sus indicadores sensoriales; existe un número indefinido de análisis perceptivos que serán, en principio, compatibles con cualquier especificación determinada de un input sensorial[18]. Según esto, la forma más plausible de entender las integraciones perceptivas es considerarlas como especies de inferencias-sobre-la-mejor-explicación, teniendo en cuenta que el problema computacional en la integración perceptiva es el de escoger la mejor hipótesis sobre la fuente distal de las estimulaciones proximales.

Existe, pues, un gran problema en cuanto a la forma de relacionar las condiciones para aplicar las descripciones físicas con las condiciones para aplicar descripciones como «la hora del té». Lo que quiero señalar aquí es que las capacidades computacionales del organismo deben constituir una solución a estos problemas en la medida en que sus juicios perceptuales sean: a) mediados por la información sensorial, y b) verdaderos.

Es ya la hora de llegar a una conclusión, que por otra parte sonará a estas alturas como algo muy conocido. Si se acepta, aunque sea a grandes rasgos, el tipo de enfoque de la percepción que acabamos de exponer, se tiene que aceptar el punto de vista de que los procesos perceptivos suponen la computación de una serie de redescripciones de los estímulos del entorno. Pero esto es reconocer que la percepción pre-

---

[18] De ahí la posibilidad de las ilusiones perceptivas. Puede verse una exposición de la percepción que está dentro de las coordenadas que estoy siguiendo en Gregory (1966) o Teuber (1960).

supone un sistema representacional; en realidad, un sistema representacional lo suficientemente rico como para distinguir entre los miembros de los conjuntos de propiedades que, en su totalidad, son manifestados por el mismo hecho. Si, por ejemplo, *e* es una instancia de un tipo de frase, y si la comprensión/análisis perceptivo de *e* exige que se determine de qué tipo de frase es instancia (véase la primera parte del Capítulo 3), entonces, y según el presente enfoque de la comprensión/análisis perceptivo, habrá que computar una serie de representaciones de *e*. Esta serie tendrá que incluir, y distinguir entre, las representaciones que especifiquen las propiedades acústicas, fonológicas, morfológicas y sintácticas de la instancia en cuestión. Deberá incluir todas estas representaciones porque, por lo que se puede saber, cada una de ellas es esencial para determinar la relación instancia/tipo en relación con las elocuciones de las frases. Tendrá que distinguir entre ellas porque, por lo que podemos saber, las propiedades de las frases que se definan según cualquiera de estas clases de representación quedarán, ipso facto, indefinidas en relación con cualesquiera otras.

Volvemos así a nuestra afirmación de que los procesos psicológicos son computacionales y la computación presupone un medio para representar las estructuras en que se definen las operaciones computacionales. En vez de insistir otra vez en este punto, acabaré esta parte de la exposición explicando dos suposiciones de las que depende el argumento.

He afirmado que los únicos modelos disponibles para la decisión, el aprendizaje de conceptos y la percepción, tratan todos ellos estos fenómenos como computacionales y por lo tanto presuponen que el organismo tiene acceso a un lenguaje en que se realizan las computaciones. Pero, naturalmente, este argumento requiere que se acepten los modelos literalmente, al menos en cuanto esquemas para la explicación de los fenómenos. En concreto, requiere aceptar que si un determinado modelo atribuye un estado a un organismo, entonces, en la medida en que aceptamos el modelo, nos vemos ontológicamente obligados a aceptar tal estado. Ahora bien, hay muchos filósofos que no son partidarios de este juego. Están dispuestos a aceptar las explicaciones computacionales de los procesos cognitivos aunque no sea más que por la ausencia de otras alternativas teóricas viables. Pero los modelos se aceptan únicamente como *façons de parler,* habiéndose aceptado previamente algún programa reduccionista.

Como hice notar en la introducción, no puedo demostrar que sea imposible mantener la fuerza de las teorías psicológicas computacionales dentro de algún marco de referencia que trate los estados mentales como disposiciones conductuales (por ejemplo). Pero creo que es justo decir que no ha habido nadie que haya logrado nunca dar una razón para que creamos que es posible, y el programa parece cada vez más desesperado en la medida en que la investigación empírica revela lo complejas que son realmente las estructuras mentales de los organismos y las interacciones de tales estructuras. Ya he insinuado que en esto hay que estar a las duras y a las maduras. Si nuestras teorías psicológicas nos llevan a admitir un lenguaje del pensamiento, lo que debemos hacer es aceptarlo con seriedad y averiguar cómo es el lenguaje del pensamiento.

Mi segunda observación es que, aunque he argumentado en favor de un lenguaje del pensamiento, lo que he mostrado en realidad ha sido, en el mejor de los casos, que existe un lenguaje de la computación, pues el pensar es algo que hacen los *orga-*

nismos. Pero las clases de procesos de datos que he estado considerando, aunque pueden ser llevados a cabo por los sistemas nerviosos de los organismos, no son probablemente, en el sentido más directo, atribuibles a los mismos organismos.

Es evidente que estamos ante un problema terriblemente difícil, el de saber qué es lo que determina si el sujeto de una acción es personal (en cuanto distinto de su cuerpo o de las partes de su cuerpo). Muchos filósofos ponen tremendo énfasis en trazar esta distinción, y no es para menos: puede ser crucial en contextos como la evaluación de la responsabilidad legal o moral. También puede resultar crucial cuando el objetivo es la fenomenología: es decir, la caracterización sistemática de los estados *conscientes* del organismo[19]. Pero cualquiera que sea la importancia que pueda tener para *ciertos* objetivos la distinción entre los estados del organismo y los estados de su sistema nervioso, no hay ninguna razón especial para suponer que es importante para los objetivos de la psicología cognitiva.

Lo que suelen tratar de hacer los psicólogos cognitivos es describir la etiología de la conducta desde el punto de vista de una serie de transformaciones de información. Véase la segunda parte del Capítulo 2, en que se explica esta noción con mayor amplitud; pero, hablando en términos aproximados, se dice que la información está a disposición del organismo cuando el hecho neural que la codifica es uno de los determinantes causales de la conducta del organismo. El término «conducta» se interpreta en sentido amplio (e intuitivo) de forma que incluye, por ejemplo, el pensar y el soñar, pero no la aceleración que se produce al caerse por las escaleras.

Si se tienen presentes estos objetivos, resulta (una vez más, por motivos empíricos más que conceptuales) que la distinción ordinaria entre lo que hace, conoce, piensa y sueña el organismo y lo que ocurre a y en su sistema nervioso, no parece tener una importancia excesiva. Parece que las clases naturales, con miras a la construcción de teorías, incluyen cosas que hace el organismo, cosas que ocurren en el sistema nervioso del organismo, y cosas que ocurren en su entorno. No parece conveniente que los filósofos insistan en que, dado que esta clase de teoría no traza las distinciones habituales, esta teoría *tiene que* ser un embrollo. No se puede presentar como objeción para una teoría el que existan distinciones que ella no hace; de lo contrario, sería una objeción que habría que hacer a todas las teorías. (Los aristotélicos consideraron como argumento *contra* la mecánica de Galileo el que no distinguiera entre cuerpos sublunares y celestes; es decir, que se aplicaran sus generalizaciones a cada uno de ellos. Esta línea de argumentación suele considerarse de forma general como poco atinada.)

---

[19] Lógicamente, no está nada claro que se pueda realizar esta tarea de forma reveladora. Esto dependerá de si *hay* generalizaciones que sean válidas (precisamente) para los estados mentales conscientes, y eso depende a su vez de si los estados conscientes de un organismo tienen más cosas en común entre sí que con los estados *in*conscientes del sistema nervioso del organismo. En este sentido sigue siendo una cuestión sin aclarar el tema de si los estados psicológicos conscientes constituyen un dominio natural para una teoría, lo mismo que sigue siendo una cuestión sin aclarar el tema de si, por ejemplo, todos los objetos de Minnesota constituyen un dominio natural para una teoría. No se pueden tener teorías de todo en todas y cada una de sus descripciones, y normalmente no se puede resolver a priori qué descripciones de qué cosas son las que se pueden generalizar. Me inclinaría a pensar que, a partir de Freud, la obligación de probar ha pasado a aquellos que mantienen que los estados conscientes (de los seres humanos) forman un dominio teórico.

En resumen, los estados del organismo que se postulan en las teorías de la cognición no deberían considerarse como estados del organismo en relación con una teoría de la responsabilidad legal o moral, por ejemplo. Lo que importa es que se puedan considerar como estados del organismo con miras a *algún* objetivo útil. En concreto, lo que importa es que se puedan considerar como estados del organismo con miras a la construcción de teorías psicológicas que sean verdaderas.

Podríamos decir lo mismo pero al revés. Si las teorías psicológicas no consiguen trazar las distinciones habituales entre algunas de las cosas que ocurren a los organismos y algunas de las cosas que hacen los organismos, eso *no* implica que los psicólogos tengan que negar que existan tales distinciones o que haya que trazarlas para algún objetivo. Tampoco implica que los psicólogos se vean obligados (de alguna manera, y cualquiera que sea la significación precisa de la palabra) a «volver a trazar la geografía lógica» de nuestros conceptos mentales corrientes. Lo que *sí* implica (y todo lo que se implica) es sencillamente que la distinción entre acciones y acontecimientos no es una distinción *psicológica*. Después de todo, hay muchas distinciones muy precisas que no lo son[20].

---

[20] Estos comentarios están en conexión evidente con los que servían de conclusión a la introducción: las distintas disciplinas intelectuales constituyen generalmente una clasificación cruzada de sus objetos materiales.

Capítulo 2

LENGUAJE PRIVADO, LENGUAJES PUBLICOS

El interior no es el exterior.

SØREN KIERKEGAARD

POR QUE TIENE QUE HABER UN LENGUAJE PRIVADO

Podríamos resumir las páginas anteriores de la siguiente manera: una de las variables esenciales en cualquier teoría de los procesos cognitivos superiores que podamos imaginar es el carácter de la representación que el organismo atribuye a los rasgos de su entorno y a sus opciones de respuesta. Es obvio que se trata de una afirmación muy tradicional. Los psicólogos de la Gestalt, por ejemplo, solían insistir en la prominencia del estímulo *proximal* en la causación de la conducta. Con ello indicaban que para saber cómo va a responder un organismo a un hecho del entorno, hay que averiguar antes cuáles son las propiedades que considera propias de tal hecho[1]. Con el mismo derecho, podrían haber resaltado el carácter prominente de la *respuesta* proximal; para saber por qué se comportó el organismo en la forma en que lo hizo, hay que averiguar antes a qué descripción trataba de responder su conducta; qué es lo que suponía que estaba haciendo. En el Capítulo 1 intentamos explicitar una de las presuposiciones de esta línea argumental: el «estímulo proximal» es una *representación* proximal del estímulo distal, y la «respuesta proximal» *representa* un acto manifiesto. Pero la representación presupone un medio de representación, y no hay simbolización si no hay símbolos. En concreto, no existe representación interna sin lenguaje interno.

Personalmente, considero que esta conclusión es verdadera y además enormemen-

[1] No sólo porque la conducta se basa algunas veces en falsas creencias (por ejemplo, en la atribución equivocada de propiedades al estímulo), sino también porque las propiedades conductualmente sobresalientes del estímulo son una *selección* de las propiedades que le pertenecen: del número indefinido de propiedades que tiene el estímulo, sólo pueden ser conductualmente sobresalientes aquellas que el organismo *representa* como presentes en el estímulo. Esa es la razón por la que, en la práctica, normalmente la única manera de saber cuál es el estímulo (proximal) es atender a la conducta del organismo.

te importante. Sin embargo, existen posibles interpretaciones que harían que fuera verdadera pero no muy importante. Por ejemplo, podría argumentarse de la siguiente manera:

Existe, lógicamente, un medio en el que pensamos, y que, lógicamente, es un lenguaje. De hecho, se trata de un lenguaje natural: el inglés para los que hablan inglés, el francés para los que hablan francés, el hindi para los que hablan hindi, etc. Así, el argumento que parecía llevar a conclusiones interesantes y paradójicas no nos lleva ni a la puerta de casa. Las «afirmaciones tradicionales» se basan, en resumidas cuentas, en una confusión tradicional. Se supone que el lenguaje natural es el medio en que *expresamos* nuestros pensamientos, cuando en realidad es el medio en que los *pensamos*.

Es esta una opinión que ha convencido a muchísimos filósofos y psicólogos. Hay que reconocer que resulta atractiva, pues con ella el teórico puede admitir el papel esencial de la computación (y por tanto de la representación) en la producción de la conducta y al mismo tiempo resistir a las implicaciones más fuertes de la idea de un lenguaje del pensamiento. Por ejemplo, no hay problemas en que la formación de hipótesis sea esencial para el aprendizaje y en que las hipótesis presupongan un lenguaje en que se formulen, con tal que el lenguaje presupuesto sea, por ejemplo, el inglés. El inglés es un sistema representacional cuya existencia debemos reconocer independientemente de nuestras opiniones sobre la psicología cognitiva; y si no, que se lo pregunten a cualquier anglófono. Es decir, podemos admitir que los procesos cognitivos están definidos en relación con los objetos lingüísticos y hacerlo sin provocar ningún tipo de recelo metodológico. Lo único que tenemos que hacer es suponer que los objetos lingüísticos en relación con los cuales se definen los procesos cognitivos están tomados de uno de los lenguajes *públicos*.

El único fallo de esta proposición es que no es posible tomársela en serio: en mi opinión, tendremos que acostumbrarnos a aceptar las consecuencias radicales del lenguaje interno. La refutación obvia (y suficiente, diría yo) de la afirmación de que los lenguajes naturales son el medio del pensamiento es que hay organismos no verbales que piensan. No tengo intención de hacer casuística sobre lo que debe considerarse como pensamiento, por lo que me aclararé haciendo referencia a los ejemplos expuestos en el Capítulo 1. Los tres procesos que examinamos entonces —acción considerada, aprendizaje de conceptos e integración perceptual— son logros bien conocidos de los organismos infrahumanos y de los niños pre-verbales. Por consiguiente, lo menos que se puede decir es lo que venimos diciendo todo el tiempo: los modelos computacionales de estos procesos son los únicos con que podemos contar. Los modelos computacionales presuponen sistemas representacionales. Pero es indudable que los sistemas representacionales de los organismos pre-verbales e infrahumanos no pueden ser lenguajes naturales. Entonces, o abandonamos la psicología pre-verbal e infrahumana que hemos estado construyendo, o admitimos que parte del pensamiento, al menos, no se hace en inglés.

Téngase en cuenta que aunque la computación presupone un lenguaje representacional, *no* presupone que ese lenguaje deba ser uno de los que funcionan como vehículos de comunicación entre hablantes y oyentes: por ejemplo, que deba ser un lenguaje natural. Por eso no hay ninguna razón interna para suponer que nuestra psicología sólo se aplique a los organismos que hablan, y si decidimos restringir de esa

manera su aplicación no tendremos ningún modelo para el aprendizaje, la elección y la percepción en poblaciones que no sean las de los seres humanos que hablan un idioma. Por otra parte, ampliar nuestra psicología a especies infrahumanas significa aceptar procesos cognitivos mediados por sistemas representacionales distintos de los lenguajes naturales.

Tengo la impresión de que muchos filósofos no se dejan impresionar por consideraciones de esta índole, ya que están convencidos de que no es cuestión de hecho sino, por así decirlo, de política lingüística el que los predicados psicológicos que tienen aplicaciones paradigmáticas a seres humanos que hablan un idioma se deban «ampliar» a lo que no pasa de ser infraverbal. Una vez un filósofo muy joven me dijo que la cuestión de si los animales pueden oír (de si se puede decir que oyen) es cuestión de *decisión*. «Después de todo», dijo, «la palabra es *nuestra*».

Pero esta forma de convencionalismo no tiene muchas posibilidades; la cuestión no es si debemos ser corteses con los animales. En concreto, existen homogeneidades entre las capacidades mentales de los organismos infraverbales y las de los seres humanos que hablan un idioma que, como es bien sabido, son inexplicables a no ser que se acepte que la psicología infraverbal es pertinentemente homogénea en relación con nuestra psicología.

Por citar sólo un ejemplo, en el Capítulo 1 indicamos que los sujetos humanos suelen tener más dificultades para dominar los conceptos disyuntivos que los conjuntivos o negativos. Pero también indicamos que había que considerar la noción de la forma de un concepto en relación con el sistema de representación que emplea el sujeto. En primer lugar, la disyunción es interdefinible con la conjunción y la negación, y, en segundo lugar, el que los conceptos sean o no disyuntivos depende de cuáles son los términos de clase que reconoce el vocabulario del sistema representacional. *El color* no es un concepto disyuntivo a pesar del hecho de que los colores sean diferentes. En cambio, «rojo o azul» *es* un concepto disyuntivo, es decir, está representado disyuntivamente en inglés y, probablemente, en cualquier sistema de representación que intervenga como mediador en la integración de nuestros perceptos visuales.

Lo importante es que estas observaciones se pueden aplicar en su totalidad al aprendizaje de conceptos infraverbal. También a los animales les cuesta dominar (lo que *nosotros* consideramos como) conceptos disyuntivos. Podemos explicar este hecho si suponemos que el sistema representacional que utilizan *ellos* es considerablemente semejante al que utilizamos *nosotros* (por ejemplo, que un animal condicionado a responder positivamente a o-un-triángulo-o-un-cuadrado representa las contingencias de reforzamiento en forma disyuntiva, tal como hace el experimentador)[2]. Como no se nos ocurre ninguna otra explicación alternativa (pues nunca, que yo sepa, se ha presentado ninguna explicación alternativa) habría que pensar que son los hechos conductuales, y no nuestras aptitudes lingüísticas, los que nos obligan a partir de la hipótesis de que existen homogeneidades pertinentes entre nuestro sistema representacional y los que utilizan los organismos infraverbales[3].

---

[2] Véase Fodor, Garrett y Brill, 1975, donde el lector puede encontrar una demostración experimental de que, en la etapa preverbal, los niños tienen dificultades diferenciales con las contingencias de refuerzo disyuntivas.

[3] Conviene insistir en que este ejemplo no tiene nada de especial. La difundida homogeneidad de los

Como era de esperar, estos problemas adquieren importancia fundamental cuando pensamos en el niño pre-verbal que aprende su primer idioma. Lo primero que hay que hacer notar es que no tenemos la más remota idea de cómo se aprende el primer idioma sin recurrir a una u otra versión del aprendizaje por formación y confirmación de hipótesis. Esto no es extraño, pues como señalamos en el Capítulo 1, prescindiendo de los casos en que lo que se aprende es algo que se enseña explícitamente, no tenemos idea de cómo se aprende *ninguna* clase de conceptos como no sea mediante la formación y confirmación de hipótesis. Y el aprender un lenguaje L debe implicar, cuando menos, aprender conceptos como «oración de L».

Por ejemplo, si Chomsky está en lo cierto (véase Chomsky, 1965; hay una exposición detallada de los puntos de vista de Chomsky sobre la adquisición de la sintaxis en Fodor et al., 1974), el aprender un primer lenguaje implica la construcción de gramáticas que estén en consonancia con un sistema de universales lingüísticos determinado innatamente y poner a prueba estas gramáticas comparándolas con un conjunto de elocuciones observadas, siguiendo un orden fijado por una métrica de simplicidad que es innata. Y, lógicamente, debe haber un lenguaje en que se representen los universales, las gramáticas opcionales y las elocuciones observadas. Y, naturalmente, este lenguaje no puede ser un lenguaje natural pues, por hipótesis, lo que está aprendiendo el niño es su primer lenguaje[4].

Sin embargo, en este contexto no importa si Chomsky tiene o no razón, pues se podría hacer la misma observación basándose en suposiciones mucho más modestas sobre lo que ocurre en la adquisición del lenguaje. Me interesa tratar este tema con cierto detalle.

En primer lugar, quiero dar tres cosas por supuestas: (1) que aprender el primer lenguaje es cuestión de formación y confirmación de hipótesis en el sentido examina-

---

procesos mentales humanos e infrahumanos ha constituido el tema principal de la teoría psicológica a partir de Darwin. Los casos interesantes, inquietantes y excepcionales son aquellos en que aparecen diferencias entre las especies. Así, por ejemplo, hay situaciones que los organismos infrahumanos tratan como homogéneos estímulos que *nosotros* consideramos como disyuntivos. Es muy difícil adiestrar al pulpo para que discrimine líneas diagonales que difieren (únicamente) en la orientación izquierda-derecha. La suposición natural es que el sistema de representaciones que emplea el animal no distingue entre (es decir, atribuye representaciones idénticas a) las imágenes especulares. Véase una explicación ingeniosa en Sutherland (1960).

[4] El argumento de Chomsky infiere el carácter innato de la información lingüística (y por eso del sistema representacional en que se formula) a partir de la universalidad de la estructura lingüística en comunidades históricamente separadas y a partir de la complejidad de la información que el niño debe dominar para llegar a hablar con fluidez. Pueden encontrarse versiones de este argumento en Katz (1966) y Vendler (1972). Creo que es un argumento válido, aunque deja en el aire algunas cuestiones. Mientras no sepamos *qué* rasgos del lenguaje son universales, no nos permite saber qué aspectos de la representación por el niño de su lenguaje nativo son innatos. Y *¿hasta qué punto* ha de ser complejo el aprendizaje para que sea plausible la hipótesis de una contribución innata, específica de la tarea en cuestión?

Las reflexiones que voy a exponer tratan de delinear aspectos de la contribución innata del niño al aprendizaje lingüístico de tal manera que se superen estas dificultades. Pero daré por supuesto lo que Chomsky et al., han dado siempre por supuesto y lo que Vendler ha formulado explícitamente: existe una analogía entre aprender un segundo lenguaje partiendo de la base del primero y aprender un primer lenguaje partiendo de la base de una dotación innata. En uno y otro caso, debe explotarse algún sistema representacional disponible de antemano para formular las generalizaciones que estructuran el sistema que se está aprendiendo. De la nada no sale nada.

do en el Capítulo 1; (2) que el aprendizaje del primer lenguaje implica al menos aprender las propiedades semánticas de sus predicados; (3) que *S* aprende las propiedades semánticas de *P* únicamente si *S* aprende alguna generalización que determine la extensión de *P* (es decir, el conjunto de cosas de las que *P* es verdad).

Estas suposiciones son desigualmente tendenciosas. La número (1) se basa en los argumentos considerados en el Capítulo 1. Considero que la número (2) la aceptarán todos los que estén dispuestos a admitir que hay algo positivo en la idea de que las propiedades semánticas son psicológicamente reales. Por otra parte, la número (3) es importante, pero no voy a argumentar en su defensa, pues, como se apreciará en seguida, se adopta principalmente en beneficio de la exposición. Baste señalar que muchos filósofos han considerado plausible la afirmación de que sólo se comprende un predicado si se saben las condiciones en que serían verdaderas las oraciones que lo contienen. Pero si esto es cierto, y si, como hemos supuesto, el aprendizaje del lenguaje es cuestión de comprobación y confirmación de hipótesis, entre las generalizaciones sobre un lenguaje que debe plantear como hipótesis y confirmar el que lo aprende hay algunas que determinan las extensiones de los predicados de ese lenguaje. Una generalización que realice esta determinación es, por previo acuerdo, una *regla de verdad.* Todo esto se puede resumir en la fórmula siguiente: «*S* aprende *P* sólo si *S* aprende una regla de verdad para *P*»[5].

Como tengo intención de explotar estas suposiciones hasta el fondo, es mejor que formule algunas advertencias cuanto antes. Son tres. En primer lugar, que, aunque para lo que me propongo es conveniente identificar el aprendizaje de las propiedades semánticas de *P* con el aprendizaje de una regla de verdad con respecto a *P,* no hay

[5] Voy a utilizar en todo momento el siguiente formato para las reglas de verdad. Donde *P* es un predicado del lenguaje que se va a aprender, *T* es una regla de verdad para *P* si y sólo si (a) es de la misma forma que *F,* y (b) todos sus *casos de sustitución son*

F: ⌜P_y⌝ es verdad (en L) si y sólo si x es G

verdaderos. Los casos de sustitución de *F* son las fórmulas obtenidas:

1. Reemplazando los ángulos por comillas. (En efecto, las variables que van entre ángulos se considera que abarcan las expresiones del lenguaje objeto.)
2. Reemplazando «P_y» por una frase cuyo predicado sea *P* y cuyo sujeto sea un nombre u otra expresión de referencia.
3. Reemplazando «*x*» por una expresión que designe el individuo a que se refiere el sujeto de la frase citada. (Esta condición da lugar a la noción no sintáctica de *caso de sustitución,* pues el que una fórmula tenga o no relación con otra dependerá, en parte, de qué es a lo que se refieren las expresiones de referencia. Sin embargo, esto es ventajoso e inofensivo para nuestros objetivos).

Así pues, supongamos que *L* es el inglés y *P* es el predicado «is a philosopher» (=«es filósofo»). En ese caso, una función de verdad plausible para *P* es ⌜*y* is a *philosopher*⌝ *es verdad si y sólo si x es filósofo.* Los casos de sustitución de esta regla de verdad incluirían «*Fred es filósofo*» *es verdad si y sólo si Fred es filósofo;* «*el hombre que está en el rincón es filósofo*» *es verdad si y sólo si el hombre que está en el rincón es filósofo;* y «*Fred es filósofo*» *es verdad si y sólo si el hombre que está en el rincón es filósofo* (suponiendo que el hombre que está en el rincón es Fred)..., etc.

Naturalmente, no hay nada que exija que la expresión que forma la parte de la derecha de una regla de verdad (o de sus casos) tenga que proceder del mismo lenguaje que la frase citada a la izquierda. Por el contrario, veremos que esta suposición es bastante *in*verosímil cuando supongamos que en el aprendizaje de una lengua interviene el aprendizaje de las reglas de verdad. (Véase en Davidson, 1967, una introducción práctica al programa general del análisis del significado en términos de verdad).

nada fundamental dentro del argumento que voy a presentar que dependa de que lo hagamos. Los lectores que se opongan a la identificación son libres de sustituirla por alguna otra noción de propiedad semántica o de considerar que tal noción no se ha analizado. En segundo lugar, decir que alguien ha aprendido una regla de verdad para un predicado no es decir que haya aprendido un procedimiento para determinar cuándo se aplica el predicado, ni siquiera que exista dicho procedimiento. En tercer lugar, si hubiera algo de verdad en las explicaciones basadas en la idea de disposición conductual para dar cuenta de lo que interviene en la comprensión de un predicado, tendríamos una alternativa a la teoría de que el aprendizaje de un predicado implica el aprendizaje de una regla. Por eso, toda la exposición se basará en la suposición de que no se puede decir nada a favor de las explicaciones disposicionales de lo que hace falta para comprender un predicado. Trataré cada uno de estos puntos con cierta amplitud antes de volver al argumento principal.

1. Muchos filósofos piensan que las condiciones de verdad representan una interpretación demasiado débil de lo que aprendemos cuando aprendemos un predicado; por ejemplo, que lo que aprendemos debe ser lo que implican las oraciones que contienen el predicado y lo que está implicado por ellas, y no lo que presuponen materialmente o está materialmente presupuesto en ellas. Siento una gran simpatía por estos puntos de vista. Pero lo que yo quiero subrayar es que los argumentos que voy a utilizar son totalmente neutrales en relación con tales puntos de vista. Es decir, estos argumentos son neutrales en relación con la controversia entre las semánticas extensionalista e intensionalista. Si el lector es extensionalista, creerá, indudablemente, que las propiedades semánticas de un predicado determinan su extensión. Si es intensionalista, lo que creerá es que las propiedades semánticas de un predicado determinan su *in*tensión y que las intensiones determinan las extensiones. En uno y otro caso cree lo que yo he querido dar por supuesto.

Podríamos formularlo de otra manera: tanto los intensionalistas como los extensionalistas afirman que las teorías semánticas emparejan los predicados del lenguaje objeto con sus paralelos metalingüísticos. Los extensionalistas afirman que la condición crítica de las expresiones emparejadas es la coextensibilidad. Para los intensionalistas la condición crítica es la equivalencia lógica o, quizá, la sinonimia. Pero si se cumple cualquiera de estas dos últimas condiciones, también se cumple la primera. Por eso, repito, la forma en que se resuelva la cuestión extensionalista/intensionalista no afecta a mis objetivos.

Sin embargo, existen filósofos que afirman no sólo que las propiedades semánticas de un predicado no determinan su *in*tensión, sino que tampoco determinan su *ex*tensión. Estos filósofos vienen a afirmar que lo que sabemos sobre los significados de los predicados determina como mucho sus extensiones *putativas,* pero el hecho de que la extensión putativa de un predicado sea de hecho su extensión *real* es algo que, en último término, está a merced de los descubrimientos empíricos.

Así, Putnam (pendiente de publicación) afirma que cuando aprendemos «oro», «gato», «agua», etc., aprendemos estereotipos socialmente aceptados de forma que resulta *razonable creer* en cosas que se conforman a los estereotipos de que satisfacen los predicados. Pero lo que es razonable creer no tiene por qué resultar verdadero a largo plazo. Quizá hubo un tiempo en que sólo se sabía que *era* agua el agua en forma líquida. Quizá más adelante llegó a descubrirse que el hielo era agua en estado

sólido. (Indudablemente esto es ontogenéticamente plausible, aunque sea un cuento de hadas histórico.) Descubrir esto sería descubrir algo sobre la extensión *real* de «agua» (es decir, que dentro de ella se incluye el hielo). Pero si el que el hielo sea agua es un descubrimiento empírico, resulta difícil ver cómo habría sido posible que el hecho de que «agua» se aplica al hielo se viera determinado, en cualquier sentido importante, por lo que se aprende cuando se aprende lo que significa «agua». Y si eso es verdad, es difícil imaginarse cómo el aprender lo que significa «agua» podría implicar aprender algo que determina la extensión de «agua» con anterioridad a estos descubrimientos. En resumen, según esta opinión, o bien las propiedades semánticas de una palabra no son lo que aprendemos cuando aprendemos la palabra, o bien las propiedades semánticas de una palabra no determinan su extensión.

No quiero intervenir en la evaluación de estas sugerencias, pues, verdaderas o falsas, en general no tienen relación con mis afirmaciones principales. Lo que voy a exponer, fundamentalmente, es que no se puede aprender un lenguaje cuyos términos expresen propiedades semánticas que no sean expresadas por los términos de un lenguaje que ya se es capaz de utilizar. Para formular este argumento es conveniente partir de que las propiedades semánticas expresadas por un predicado son aquellas que determinan su extensión, pues, cualesquiera que puedan ser sus fallos, esa suposición atribuye cuando menos un agudo sentido de la identidad de las propiedades semánticas (dos predicados tienen las mismas propiedades semánticas si se aplican al mismo conjunto de cosas). Sin embargo, si no se confirma dicha suposición, se puede elaborar el mismo tipo de argumento dada cualquier otra noción de propiedad semántica, con tal que se mantenga que lo que se aprende cuando se aprende una palabra son sus propiedades semánticas. Por ejemplo, si lo que se aprende cuando se aprende *P* es (únicamente) que sería razonable creer que *P* se aplica si y sólo si *S*, en ese caso y según mi argumento, para aprender el lenguaje que contiene *P* hay que ser ya capaces de utilizar algún (otro) lenguaje que contenga algún (otro) término del que sería razonable creer que se aplicaría si y sólo si fuera razonable creer que se aplica *P*. Lo mismo habría que decir, *mutatis mutandis*, en relación con otras interpretaciones de la *propiedad semántica*.

Luego pasaré a hacer lo que es más conveniente: considerar que la extensión de un predicado es lo que determinan fundamentalmente sus propiedades semánticas. Pero sólo si se entiende que es posible introducir ad lib. interpretaciones alternativas de la «propiedad semántica».

2. Mantener la opinión de que aprender un predicado implica aprender una generalización que determine su extensión no equivale a estar de acuerdo con cualquier especie de verificacionismo, auque en algunas obras se aprecia cierta tendencia a confundir las dos doctrinas.

Pensemos en el predicado inglés «is a chair» (=es una silla). La opinión actual es, más o menos, que nadie ha dominado ese predicado a no ser que haya aprendido que cae dentro de una generalización como ⌜*y is a chair*⌝ es verdadero si y sólo si *Gx*. (Puede verse una referencia a la notación en la nota 5, más arriba.) Pero de ahí no se deduce que alguien que sepa lo que significa «is a chair» domine por eso mismo un procedimiento general para clasificar los estímulos en sillas y no sillas. Sólo se deduc*iría* si se supusiera, además, que tiene un procedimiento general para clasificar los estímulos en estímulos que satisfacen *G* y estímulos que no satisfacen *G*. Pero di-

cha suposición no forma parte de la opinión de que aprender un lenguaje implica aprender las reglas de verdad de sus predicados.

Si, por ejemplo, es cierto que «chair» (=silla) significa «asiento portátil individual», se puede admitir que nadie ha dominado «is a chair» a no ser que haya aprendido que cae dentro de la regla de verdad «⌜y is a chair⌝ es verdad si y sólo si x es un asiento portátil individual». Pero podría ocurrir perfectamente que alguien supiera esto de «is a chair» y que, sin embargo, no fuera capaz de decir de un objeto determinado (o de cualquier objeto) si es o no es una silla. Se encontraría en esta situación si, por ejemplo, su forma de saber si algo es una silla es tratando de descubrir si cumple con la parte de la derecha de la regla de verdad, no siendo capaz de saber si *esta* (o cualquier) cosa es un asiento portátil individual.

Hago estas precisiones teniendo en cuenta la observación de Wittgenstein de que muchos predicados lingüísticos corrientes (quizás todos) son de textura abierta; por ejemplo, que hay un número indefinido de objetos de los que no podemos decir si son sillas; no precisamente porque la iluminación sea mala o porque no se conozcan todos los datos, sino porque «is a chair» es, por así decirlo, indefinido en relación con los objetos de esas clases, de manera que el que sean sillas o no es algo que no depende en absoluto de los hechos (cf. la silla (sic) hecha con burbujas del jabón; el cajón de embalaje que se utiliza como silla, etc.). Todo esto es cierto, pero lo que estamos intentando decir ahora es que no prejuzga la idea de que aprender las reglas de verdad sea esencial al aprendizaje lingüístico, o de que las reglas de verdad se expresan mediante fórmulas bicondicionales. Todo lo que demuestra es que *si* la condición de verdad de «is a chair» se expresa mediante «es un asiento portátil individual», en ese caso «asiento portátil individual» debe tener una textura abierta, indefinida, etc., solamente en los casos donde lo sea «is a chair».

La confusión en este punto podría ser origen de problemas sin número. Por ejemplo, Dreyfus (1972), si es que le entiendo correctamente, parece adoptar la siguiente argumentación contra la posibilidad de recurrir a modelos de las capacidades lingüísticas humanas basados en las analogías con las máquinas: (a) Los modelos basados en las máquinas deberían utilizar reglas para expresar las extensiones de los predicados que utilizan. (b) Tales reglas serían probablemente bicondicionales (por ejemplo, reglas de verdad). Pero (c) Wittgenstein ha demostrado que la extensión de los predicados del lenguaje natural no se puede expresar mediante estas reglas, pues se trata de predicados que son por naturaleza de límites poco precisos. Por eso (d) las personas no pueden ser reproducidas por las máquinas y (e) a fortiori, las personas no pueden *ser* máquinas.

Pero Wittgenstein no demostró nada de eso. Lo más que se puede deducir de la existencia de una textura abierta es que si una fórmula expresa las condiciones de verdad de *P*, su valor de verdad debe ser indeterminado siempre que el valor de verdad de *P* sea indeterminado. O, dicho de forma algo diferente, si una máquina simula el uso de un predicado por un hablante, en ese caso (la máquina debería ser incapaz de determinar si el predicado se aplica) si y sólo si (el hablante es incapaz de determinar si el predicado se aplica). Pero no hay absolutamente nada en la noción de máquina, en cuanto instrumentos que siguen reglas, que nos haga pensar que no se puede cumplir esta condición. Por consiguiente, no hay nada en la noción de que el uso del lenguaje por las personas esté dirigido por reglas, que nos haga pensar que

todo predicado de un lenguaje deba tener un aplicabilidad determinada a cada uno de los objetos de predicación.

3. He dado por supuesto no sólo que aprender un predicado implica aprender algo que determina su extensión, sino también que «aprender algo que determina la extensión de *P*» debería analizarse como aprender que *P* cae dentro de una regla determinada (es decir, una regla de verdad). Ahora bien, alguien podría aceptar la primera suposición y rechazar la segunda: por ejemplo, postulando una especie de análisis conductual de «*S* sabe la extensión de *P*». O, lo que sería equivalente para estos fines, podría aceptar ambas suposiciones y postular un análisis disposicional de lo que es conocer una regla. Así, si la regla de verdad para *P* es «⌜*Py*⌝ es verdad si y sólo si *Gx*», entonces saber la regla de verdad se podría equiparar a estar dispuestos a decir *P* solamente en los casos en que se aplica *G*. De la misma manera, aprender las condiciones de verdad de *P* sería cuestión (no de hipotetizar y confirmar que se aplica la correspondiente regla de verdad, sino únicamente) de configurar adecuadamente las propias disposiciones de respuesta.

Varios filósofos que deberían estar mejor informados aceptan, aparentemente, esos puntos de vista. Sin embargo, no me voy a tomar la molestia de repetir las objeciones clásicas, pues me parece que si hay *algo* claro es que el entender una palabra (predicado, oración, lenguaje) no es cuestión de cómo se comporta uno o de cómo está dispuesto a comportarse. La conducta, y la disposición conductual, están determinadas por las interacciones de una variedad de variables psicológicas (lo que uno cree, lo que uno quiere, lo que recuerda, lo que atiende, etc.). Por eso, en general, cualquier conducta es compatible con la comprensión, o ausencia de comprensión, de cualquier predicado. Si me pagan bien soy capaz de ponerme a hacer el pino cuando alguien diga «silla». Pero, a pesar de ello, sigo sabiendo lo que significa «es una silla».

Hasta aquí he llegado con mis advertencias. Ahora quiero sacar una conclusión. El aprendizaje de una lengua (incluyendo, naturalmente, la lengua materna) implica aprender qué significan los predicados de esa lengua. Aprender lo que significan los predicados de una lengua implica aprender a determinar la extensión de estos predicados. Aprender a determinar la extensión de los predicados implica aprender que caen dentro de ciertas reglas (es decir, reglas de verdad). Pero no se puede aprender que *P* cae dentro de *R* a no ser que se tenga un lenguaje en que se puedan representar *P* y *R*. Por eso, no se puede aprender una lengua a no ser que se tenga ya un determinado lenguaje. En concreto, no se puede aprender la primera lengua a no ser que se tenga ya un sistema capaz de representar los predicados de esa lengua *y sus extensiones*. Y si no queremos caer en un círculo vicioso, ese sistema no puede ser la lengua que se está aprendiendo. Pero la primera lengua se aprende. Por eso, existen al menos algunas operaciones cognitivas que se realizan en lenguajes distintos de los lenguajes naturales.

Wittgenstein, comentando los puntos de vista de Agustín, dice:

> Agustín describe el aprendizaje de las lenguas humanas como si el niño llegara a un país extranjero y no entendiera el idioma del país[6]; es decir, como si ya tuviera un lenguaje, sólo que

[6] Por ejemplo, Agustín representa al niño como si tratara de adivinar a qué se refieren los adultos

no era precisamente ése. O también como si el niño fuera ya capaz de *pensar,* sólo que todavía no podría hablar. Y «pensar» significaría aquí algo parecido a «hablar consigo mismo» (1953, par. 32).

Wittgenstein parece dar por supuesto que semejante punto de vista es absurdo a todas luces. Pero el argumento que acabo de esbozar indica, por el contrario, que Agustín estaba en lo cierto, como se puede demostrar, y que el reconocerlo así es el primer prerrequisito de todo intento serio por comprender cómo se aprende la primera lengua.

En realidad, creo que esta forma de argumentación se puede ampliar a aspectos que tienen profundas consecuencias para casi todas las áreas de la psicología de la cognición. En la tercera parte de este capítulo presentaré algunas razones que justifican la veracidad de esta afirmación. Sin embargo, por el momento debo comenzar una disgresión relativamente larga. Quiero ocuparme de varias clases interrelacionadas de objeciones que tratan de demostrar que, por muy plausibles que puedan parecer los sucesivos pasos de dicha argumentación, éstos *tienen* que estar equivocados ya que las conclusiones a que dan lugar son incoherentes. Voy a tomarme en serio estas objeciones no sólo porque, por lo que yo sé, muchos filósofos afirman que alguna de ellas es correcta, sino también porque el hecho de ver lo que hay de erróneo en ellas puede ayudarnos a descubrir muchos de los fundamentos filosóficos de los planteamientos computacionales de la psicología. Quiero explicar cómo funcionan, en las teorías psicológicas, los recursos a las representaciones internas, porque quiero demostrar que está perfectamente justificado que tales recursos funcionen tal como lo hacen.

COMO PODRIA DARSE UN LENGUAJE PRIVADO

La primera objeción que quiero considerar es la alegación de retroceso infinito. Se puede resolver rápidamente (para una exposición más exhaustiva, véase el intercambio entre Harman, 1969, y Chomsky, 1969).

Alguien podría decir: «Según lo que dice, no se puede aprender un lenguaje a no ser que se sepa ya un lenguaje. Pero pensemos ahora en *ese* lenguaje, el metalenguaje en que se formulan las representaciones de las extensiones de los predicados del lenguaje objeto. Indudablemente, para aprender*lo* hace falta un conocimiento previo de un meta-metalenguaje en que se formulen sus definiciones de la verdad. Y así sucesivamente, hasta el infinito. Y eso es improcedente». Creo que hay una respuesta breve y definitiva. Mi opinión es que no se puede aprender un lenguaje a no ser que ya se *sepa* uno. No es que no se pueda aprender un lenguaje a no ser que se haya *aprendido* ya uno. Esto último supondría un retroceso hacia el infinito, pero no ocurre eso en el primer caso; no, al menos siguiendo la ruta que se está explorando aquí. Lo que ha demostrado realmente la objeción es que *o* mis puntos de vista son falsos *o* al menos uno de los lenguajes que uno sabe no se ha aprendido. No me supone ningún aprieto este dilema pues la segunda opción me parece completamente plausible:

---

cuando utilizan las expresiones de referencias de su lenguaje. Según la opinión de Wittgenstein esta descripción sólo tendría sentido si se parte del supuesto de que el niño tiene acceso a un sistema lingüístico en que se realiza esa «adivinación».

el lenguaje del pensamiento se sabe (por ejemplo, es el medio para las computaciones que están en la base de los procesos cognitivos) pero no se aprende. Es decir, es innato. (Compárese Atherton y Schwartz, 1974, donde se incurre explícitamente en el falso argumento que acabamos de rechazar).

Existe, sin embargo, una forma de expresar el argumento del retroceso infinito de forma más perspicaz: «Usted dice que la comprensión de un predicado implica la representación de la extensión de tal predicado en algún lenguaje que ya se entiende. ¿No presupone eso una representación de *sus* condiciones de verdad en algún meta-metalenguaje previamente entendido? Y así hasta el infinito». Este argumento se distingue del primero en que el retroceso se fija en el «entender» más que en el «aprender», y esa diferencia es importante. Aunque no acepto la afirmación de que el lenguaje del pensamiento sea *aprendido,* acepto la afirmación de que, en cierto sentido, se entiende: por ejemplo, que está disponible para su utilización como vehículo de los procesos cognitivos. Sin embargo, esta objeción, como la otra, incurre en la falacia de *ignoratio elenchi:* la posición atacada no es la que se defiende.

Lo que yo he dicho ha sido que aprender lo que significa un predicado implica representar la extensión de ese predicado; no que lo implique la comprensión del predicado. Una condición suficiente para esto último podría ser sencillamente que la utilización del predicado se pueda conformar siempre, de hecho, con la regla de verdad. Para ver lo que entra aquí en juego, podemos considerar el caso de los ordenadores.

Los ordenadores suelen utilizar al menos dos lenguajes diferentes: un lenguaje de input/output en el que se comunican con su entorno y un lenguaje de máquina en que hablan consigo mismos (es decir, en el que realizan sus computaciones). Los «compiladores» median entre los dos lenguajes, especificando bicondicionales cuya parte de la izquierda es una fórmula en el código de input/output y cuya parte de la derecha es una fórmula en el código de máquina. Estas bicondicionales son, en todos los sentidos y para todos los objetivos, representaciones de condiciones de verdad para fórmulas del lenguaje de input/output, y la capacidad de la máquina para utilizar ese lenguaje depende de la disponibilidad de esas definiciones. (Todo esto está muy idealizado, pero es lo suficientemente exacto como para servir a lo que estamos considerando)[7]. Lo que quiero dejar claro es que, aunque la máquina deba tener un compilador para utilizar el lenguaje de input/output, no necesita tener *también* un compilador para el lenguaje de la máquina. Lo que evita que se produzca un retroceso hasta el infinito en el caso de los compiladores es el hecho de que la máquina está *construida* para utilizar el lenguaje de máquina. Más o menos, el lenguaje de la máquina se distingue del lenguaje de input/output en que sus fórmulas corresponden di-

---

[7] Alguien podría señalar que, si las fórmulas del compilador son bicondicionales, se podrían considerar como especificación de las condiciones de verdad para las fórmulas del *lenguaje de máquina* donde el código de input/output suministraría los vehículos metalingüísticos de representación. Sin embargo, de hecho, la apariencia de simetría es falsa aun cuando las dos lenguas sean totalmente intertraducibles. Aunque la máquina utilice las fórmulas del código de la misma sin recurrir al compilador, no tiene acceso a fórmulas del lenguaje input/output a no ser a través de las traducciones que realiza el compilador. Por eso, hay un sentido práctico en el que, por lo que se refiere a la máquina, las fórmulas del lenguaje de máquina expresan los significados de las fórmulas en el código de input/output, pero no viceversa. Este punto se relaciona con otro que aparecerá en el Capítulo 3: los filósofos han tenido demasiada propensión a suponer que las teorías del significado están inevitablemente contagiadas de simetría.

rectamente a estados y operaciones de la máquina de carácter físico y computacionalmente pertinente: la física de la máquina garantiza así que las secuencias de estados y operaciones que realiza a lo largo de sus computaciones respeten las limitaciones semánticas de las fórmulas de su lenguaje interno. Lo que ocupa el lugar de una definición de la verdad en el caso del lenguaje de la máquina es sencillamente los principios de ingeniería que garantizan esta correspondencia.

En seguida volveré a tratar este punto con cierto detalle. Por el momento, baste con señalar que son dos las formas en que puede ocurrir que un mecanismo (incluyendo a las personas) entienda un predicado. En uno de los casos, el mecanismo tiene y emplea una representación de la extensión del predicado, donde la misma representación se da en un lenguaje que entiende el mecanismo. En el segundo caso, el mecanismo está construido de tal manera que su utilización del predicado (por ejemplo, en las computaciones) concuerde con las condiciones que especificaría dicha representación. Quiero señalar que la primera forma es cierta en el caso de los predicados de los lenguajes naturales que aprenden las personas y la segunda en el de los predicados que se dan en el lenguaje interno en que piensan.

«Pero entonces», podría replicar el lector, «se admite que hay al menos un lenguaje cuyos predicados comprendemos sin la representación interna de las condiciones de verdad. Se admite que, para ese lenguaje, la respuesta a: "¿Cómo utilizamos sus predicados correctamente?" es que lo hacemos, y nada más; que estamos hechos así. Con esto se evita la regresión hasta el infinito, pero se insinúa que también es innecesario el retroceso desde el lenguaje natural hasta el lenguaje interno. Se argumenta que aprendemos "es una silla" únicamente si aprendemos que cae dentro de la regla de verdad ⌜*y es una silla*⌝ *es verdad si y sólo si x is G* y luego se dice que no se plantea la cuestión de aprender una función de verdad para *G*. ¿Por qué no nos paramos un paso antes y nos ahorramos las complicaciones? ¿Por qué no decir que tampoco se plantea la cuestión de cómo aprendemos "es una silla"? La explicación tiene que detenerse en algún punto».

La respuesta es que la explicación tiene que detenerse en algún punto pero que no hace falta —ni es conveniente— que sea *aquí*. La cuestión de cómo aprendemos «is a chair» se plantea precisamente porque el inglés se aprende. La cuestión de cómo se aprende *G* no se plantea precisamente porque, por hipótesis, el lenguaje en que *G* es una fórmula es un lenguaje innato. También en este caso puede ser revelador pensar en los ordenadores.

La propiedad crítica del lenguaje de máquina de los ordenadores es que sus fórmulas se pueden emparejar directamente con los estados físicos computacionalmente pertinentes de la máquina, de tal manera que las operaciones que realiza la máquina respetan las limitaciones semánticas de las fórmulas del código de la máquina. Cada uno de los estados concretos de la máquina se pueden interpretar, en este sentido, como muestras de las fórmulas. Esta correspondencia se puede conseguir *también* entre los estados físicos de la máquina y las fórmulas del código de input/output, pero únicamente compilando antes estas fórmulas: es decir, únicamente si se traducen antes al lenguaje de la máquina. Esto expresa el sentido en que las máquinas *están* «preparadas intrínsecamente para utilizar» su lenguaje de máquina y *no* están «preparadas para utilizar» sus códigos de input/output. También nos hace pensar en una teoría empírica: cuando se encuentre un mecanismo que utilice un lenguaje para el

que no estaba intrínsecamente preparado (por ejemplo, un lenguaje que ha *aprendido*), hay que suponer que lo hace traduciendo las fórmulas de ese lenguaje a las fórmulas que corresponden directamente a sus estados físicos computacionalmente pertinentes. Esto se aplicaría, en especial, a las fórmulas de los lenguajes naturales que aprenden los hablantes/oyentes, y la suposición correlativa sería que las reglas de verdad para los predicados del lenguaje natural funcionan como parte del procedimiento de traducción.

Por supuesto que esto no es más que una *teoría* sobre lo que ocurre cuando alguien entiende una frase en un lenguaje que ha aprendido. Pero al menos *es* una teoría, y una teoría que hace que la comprensión de una frase sea análoga a procesos computacionales fáciles de entender en términos generales. Según este punto de vista, lo que ocurre cuando una persona entiende una frase debe ser un proceso de traducción básicamente análogo al que ocurre cuando una máquina «entiende» (es decir, compila) una frase de su lenguaje de programación. En el Capítulo 3 intentaré demostrar que hay amplias bases empíricas para considerar este tipo de modelo con la seriedad que se merece. Sin embargo, lo que ahora intento decir es sencillamente que es al menos *imaginable* que haya mecanismos que necesiten definiciones de la verdad para los lenguajes que hablan pero no para el lenguaje en que computan. Si los mecanismos somos *nosotros,* tiene sentido afirmar que aprender inglés implica aprender que ⌜y is a chair⌝ es verdad si y sólo si x es G, aun cuando se niegue que para aprender eso haga falta aprender que ⌜y es G⌝ es verdad si y sólo x es ψ para cualquier ψ que no sea G o «is a chair».

En resumidas cuentas, no creo que la visión del aprendizaje lingüístico esbozada hasta aquí lleve al regreso hasta el infinito. A lo que lleva es a retroceder un paso, es decir, del lenguaje natural al código interno —y ese paso está más motivado empírica que conceptualmente—. Es decir, podemos imaginar un organismo que nazca hablando y que nazca hablando el lenguaje que utilice su sistema nervioso para computar. En el caso de un organismo como éste, no se plantea, *ex hypothesi,* la cuestión de cómo aprende su lenguaje; y podría resultar totalmente innecesaria la idea de que su uso del lenguaje está controlado por una representación interna de las condiciones de verdad en relación con los predicados de dicho lenguaje. Lo único que tendríamos que suponer es que el organismo está construido de tal manera que su utilización de las expresiones del lenguaje se ajusta a las condiciones que articularían una definición de la verdad para ese lenguaje. Pero nosotros no somos esa clase de organismo y, por lo que yo sé, en nuestro caso no parece defendible ninguna alternativa a la idea de que sí que aprendemos las reglas que gobiernan las propiedades de las expresiones de nuestro lenguaje.

Consideraré ahora un último tipo de objeción que podría plantearse contra la coherencia conceptual de las suposiciones sobre el aprendizaje lingüístico que vengo haciendo. Al mismo tiempo que examinemos esta objeción, trataré de dejar bien claro cómo interviene el recurso a las representaciones internas en las teorías psicológicas que suponen que las representaciones internas son el medio de los procesos cognitivos. Después, volveré a ocuparme del tema principal y a considerar algunas de las implicaciones generales de la presente forma de ver el aprendizaje lingüístico en la medida en que tiene relación con la cuestión de cómo deben ser las representaciones internas.

Una forma de describir mi punto de vista es que los organismos (o, en cualquier caso, los organismos que tengan algún tipo de comportamiento) tienen no sólo los lenguajes naturales que puedan tener, sino también un lenguaje privado en que realizan las computaciones que están en la base de su conducta. Creo que eso es una justa descripción de lo que vengo diciendo, pero reconozco que algunos filósofos considerarían que se trata de un argumento de «reductio ad absurdum». Se da por supuesto que Wittgenstein ha demostrado que no puede haber eso que se llama lenguaje privado (1953, hacia la p. 258).

No es intención mía adentrarme en el marasmo de la discusión exegética que se da en torno al argumento del lenguaje privado. Lo que haré es presentar una breve reconstrucción y hacer ver que el argumento, interpretado de esa manera, no va contra los puntos de vista que he estado defendiendo. Sigue en pie el tema de si dicho argumento podría ir contra alguna *otra* reconstrucción. Pero conviene señalar que, fuera lo que fuera lo que demostró Wittgenstein, no pudo ser que es imposible que un lenguaje sea privado en el sentido en que lo es el lenguaje de máquina de un computador, porque los ordenadores *existen,* y lo que es real es posible. Insisto en esto porque, según vayamos avanzando, voy a apoyarme cada vez más en la analogía de la máquina tanto como prueba de existencia de mecanismos que no hablan en el lenguaje en que computan, como para proponer modelos empírico potenciales para la relación entre lenguajes naturales y el lenguaje del pensamiento.

Supongo que lo que a Wittgenstein le interesa fundamentalmente es demostrar que no corresponde ningún sentido determinado a la noción de que un término de un lenguaje privado se utilice coherentemente (en oposición, por ejemplo, a que se utilice al azar). A este respecto, Wittgenstein tiene dos formas de caracterizar un lenguaje privado: o como aquel cuyos términos se refieren a cosas que sólo puede experimentar el que lo habla o como lenguaje para la aplicabilidad de cuyos términos no existen criterios públicos (o normas o convenciones). Para lo que se proponía Wittgenstein (que debía ser fundamentalmente atacar la idea de un lenguaje de datos sensoriales) estas formulaciones vienen a reducirse prácticamente a lo mismo: si yo soy el único que sabe a qué se refiere un término como «cosquilleo ligero», en ese caso es evidente que no pueden ser públicas las convenciones para la aplicación de dicho término. Por hipótesis, sólo yo podría decir cuándo se cumplen esas convenciones; sólo yo sabría si un hecho determinado es de la clase que entra dentro de las convenciones.

Pero, según Wittgenstein, tampoco yo lo sabría. Supongamos que creo que un hecho determinado (la presencia de una sensación mía) es del tipo que se puede describir adecuadamente como tener un ligero cosquilleo. En ese caso caben dos posibilidades: o hay algo —alguna prueba— que pueda demostrar que tengo razón al utilizar este término para describir este tipo de hecho o no lo hay. Supongamos que se da la prueba. Entonces, si yo puedo recurrir a ella, ¿por qué no pueden hacerlo los demás? Es decir, si existe dicha prueba, es de suponer que es de propiedad pública, al menos en principio. Pero si hay razones públicas para creer que se pueden aplicar los términos de mi lenguaje, ya no se trata, por definición, de un lenguaje privado[8].

[8] La expresión inglesa «mild tickle» (= cosquilleo ligero) es, naturalmente, paradigma de un término de lenguaje *público;* en concreto, hay muchas formas de saber si lo estoy utilizando mal, y estas formas

Consideremos entonces la otra posibilidad: que no hay *nada* que demuestre que «cosquilleo ligero» se aplica propiamente a sensaciones como la que estoy experimentando. Si no hay pruebas de ello, no hay ninguna diferencia entre utilizar el término de forma correcta o incorrecta: ninguna diferencia entre obedecer las convenciones sobre el uso del término y el no obedecerlas. Pero una convención tal que su acatamiento o su no acatamiento vengan a ser lo mismo no es en absoluto una convención. Y un término que no está dirigido por una convención es un término que se puede utilizar al azar. Y un término que se puede utilizar al azar no es en absoluto un término. Y un lenguaje sin términos no es en absoluto un lenguaje. Pero si no es un lenguaje, a fortiori, no es un lenguaje privado.

Ahora bien, un sistema representacional interno del tipo que yo he hipotetizado, sería un lenguaje privado según el segundo criterio, aun cuando no lo fuera según el primero. Es decir, es innegablemente cierto que la aplicabilidad de los términos en el supuesto lenguaje del pensamiento no está determinado por convenciones públicas, aunque no haya ninguna razón especial que nos haga suponer que aquello a lo que se aplican los términos tengan que ser hechos privados; podrían aplicarse a números, o sillas, o predicados de inglés, o a gente pelirroja, etc. En resumen, aunque no hay nada que exija que el lenguaje del pensamiento deba interpretarse como un lenguaje de datos sensoriales, puede parecer, sin embargo, que cae dentro del ámbito del argumento de Wittgenstein y por lo tanto que corre el peligro de que ese argumento sea válido. ¿Qué hay que hacer al respecto?

En primer lugar, parece claro que el argumento del lenguaje privado no está dirigido realmente contra la clase de teoría que he estado defendiendo. No hay ninguna razón que obligue a un mentalista a suponer que las operaciones mentales tienen una intimidad epistémica en ningún sentido fuerte de tal noción. En realidad, haría bien en no suponerlo si quiere que sus teorías psicológicas sean compatibles con la ontología materialista; los fenómenos neurológicos son públicos.

Supongo que Wittgenstein podría argumentar que los testimonios neurológicos en favor del uso coherente de los términos del lenguaje interno serían irrelevantes incluso en el caso de que existieran. De hecho, nosotros no utilizamos criterios neurológicos para determinar que alguien ha dominado el uso de un término cuando, por ejemplo, le estamos enseñando una lengua. Pero esto no tendría nada que ver, y además en un doble sentido. En primer lugar, el lenguaje del pensamiento se supone que es innato. Por eso, aunque haya obligación de dar sentido a la noción de su uso coherente, no hay ninguna obligación de demostrar cómo se podría enseñar o aprender. En segundo lugar, los testimonios de que el lenguaje del pensamiento se usa coherentemente podrían ser empíricos sin ser neurológicos. Por ejemplo, podrían tener la condición de ser la mejor explicación existente sobre la coherencia global de la vida mental de los organismos.

---

son igualmente asequibles a otras personas que no son yo. Supongamos el caso de un extranjero que aprende inglés y que se plantea la cuestión de si no ha interpretado mal la frase «mild tickle». Imaginemos, por ejemplo, que parece posible que considere que «mild tickle» significa lo que significa en realidad la expresión «green afterimage» (=postimagen verde). Lo que dice Wittgenstein es que no habría ningún problema filosófico para que él (o nosotros) lo descubriera. Lo cual indica que la ausencia de problema *filosófico* equivale claramente a la ausencia de problema *práctico*.

Además, hay que decir que el argumento del lenguaje privado —al menos tal como yo lo vengo interpretando— no es demasiado bueno. Como han señalado muchos filósofos, lo más que demuestra el argumento es que a no ser que haya procedimientos públicos para *aclarar* si un término se aplica coherentemente, no puede haber forma de *saber* si se aplica coherentemente. Pero de ahí no se deduce que no *habría* de hecho diferencia entre aplicar el término coherentemente y aplicarlo al azar. A fortiori, tampoco se deduce que no tenga ningún *sentido* afirmar que hay una diferencia entre aplicar el término coherentemente y aplicarlo al azar. Quizá podrían deducirse estas consecuencias partiendo del principio verificacionista de que una afirmación no puede ser sensata a no ser que haya alguna forma de saber si es verdadera, pero es indudable que no tenemos nada que decir en favor de ese principio.

Téngase en cuenta (y éste es, para nosotros, el punto crucial) que el uso de un lenguaje para la computación no exige que uno sea capaz de *determinar* que sus términos se emplean coherentemente; sólo hace falta que su utilización *sea* de hecho coherente. Alguien podría presentar la siguiente objeción: «Imaginemos una persona que está haciendo sumas, y supongamos que no tiene ningún sistema que le permita cerciorarse de que el número para referirse al cual estuvo utilizando hace cinco minutos el numeral «2» es el mismo número para referirse al cual está utilizando ahora el numeral «2». En ese caso, indudablemente, *no podría* utilizar los numerales para realizar sus cálculos». Pero lo que es indudable es que podría hacerlo. La validez de las deducciones no se impugna con la *posibilidad* de equivocación, sino únicamente con ejemplos de equivaciones reales. Naturalmente, si el pobre hombre llegó a convencerse (por ejemplo leyendo filosofía barata) de que podría estar utilizando de hecho los numerales al azar, la *fe* en sus cálculos vacilaría considerablemente. Sin embargo, si hay un lenguaje del pensamiento, su utilización no se apoya en la fe. Lo utilizamos en la forma en que lo hacemos no por convencimiento filosófico sino por necesidad biológica.

Sin embargo, una cosa es acusar a Wittgenstein de verificacionismo, y otra muy distinta aceptar el desafío que propone el argumento del lenguaje privado. Debemos dar sentido a la idea de que los términos de un sistema representacional interno se utilizan coherentemente y demostrar cómo ese sentido es al menos razonablemente análogo al sentido en que se pueden utilizar coherentemente los términos de los lenguajes públicos. Si no conseguimos hacer lo primero, quizás es que la misma noción de lenguaje del pensamiento no sea coherente. Si no conseguimos lo segundo, no tiene mucho sentido llamar *lenguaje* al lenguaje del pensamiento.

Creo que Wittgenstein tiene una cierta idea de lo que viene a ser la coherencia de utilización en relación con los términos de un lenguaje *público* (por ejemplo, el inglés). Dicho de forma muy aproximada, el uso de los términos del lenguaje público está controlado por las convenciones de la comunidad de habla. Estas convenciones relacionan los términos (de formas muy diferentes) con situaciones públicas paradigmáticas. Utilizar coherentemente un término es utilizarlo de acuerdo con las convenciones dominantes. Utilizarlo de acuerdo con las convenciones dominantes es utilizarlo cuando se cumplen los paradigmas. En pocas palabras, un término se emplea coherentemente cuando su uso está controlado (en las formas adecuadas) por los datos sobre el mundo.

Ahora bien, lo primero que hay que observar es que—al margen de las preocupa-

ciones por los lenguajes públicos frente a lenguajes privados—esta imagen no puede ser correcta. Supongamos que tengo las mejores intenciones: supongamos, en el caso extremo, que trato de utilizar un término en aquellas situaciones, y sólo en aquellas situaciones, que son paradigmáticas para tal término. Sin embargo, mis verbalizaciones están determinadas no solamente por mis *intenciones* sino también por mis creencias. Por eso, más en concreto, el grado de correspondencia que puedo conseguir realmente entre mi uso de *P* y la ocurrencia de situaciones-*P* paradigmáticas depende no sólo de mi actitud lingüística con relación a *P* sino también de mi habilidad para determinar qué situaciones *son* situaciones-*P*. Si mis creencias están muchas veces totalmente descaminadas, es posible que haya muy poca o ninguna correspondencia entre lo que digo y la forma de ser del mundo. Pero, a pesar de todo, puede ser cierto que haya cierto sentido en la noción de que los términos de mi lenguaje están utilizados coherentemente. *P* puede ser el término que se aplica, paradigmáticamente, en las situaciones-*P* aun cuando yo no lo aplique, y no lo aplique de forma habitual.

Lo importante es que, aun en el caso de los lenguajes públicos, la coherencia no exige una relación estable entre cómo se utilizan los términos y cómo es el mundo: lo que hace falta es una relación estable entre cómo se utilizan los términos y *cómo cree el hablante que es el mundo*[9]. Es decir, lo que *sí* parece ser esencial para la utilización coherente de un lenguaje es la existencia de una cierta correspondencia entre las actitudes proposicionales y las prácticas lingüísticas del hablante/oyente; en concreto, entre lo que cree que son los hechos y las formas de palabras que considera verdaderas. Por eso, en una primera aproximación (Smith utiliza «Jones está enfermo» para representar el estado de cosas en que Jones está enfermo) si y sólo si (Smith asiente a afirmaciones hechas utilizando la *forma de palabras* «Jones está enfermo» si y sólo si Smith cree que Jones está enfermo)[10]. De la misma manera (Bill utiliza «Morris es lingüista» para representar el estado de cosas en que Morris es lingüista) si y sólo si (Bill asiente a afirmaciones hechas utilizando la forma de palabras «Morris es lingüista» si y sólo si Bill cree que Morris es lingüista). Y, en general (*S* utiliza ⌜*a* es *P*⌝ para representar el estado de cosas en que *a* es *F*) si y sólo si (*S* asiente a las afirmaciones hechas utilizando la forma de palabras ⌜*a* es *F*⌝ si y sólo si *x* cree que *a* es *F*)[11].

[9] La comunicación entre el hablante y el oyente exige, hablando en términos aproximativos, que el oyente sea capaz de decidir *lo que cree el hablante* partiendo de lo que dice el hablante (véase Capítulo 3). Cuando las creencias del hablante son *verdaderas*, el oyente será también capaz de deducir cómo es el mundo a partir de lo que dice el hablante. Quizá *para* esto sea para lo que sirve la comunicación, pero no es necesario para que se produzca comunicación.

[10] En esto se incluye, naturalmente, el asentir a sus propias afirmaciones. Por cierto, no estoy presuponiendo que asentir sea una forma de *conducta*, por lo que el presente análisis no pretende ser reductivo.

[11] Esto no es verdad, evidentemente. En primer lugar, *x* puede tener *muchas* formas de representar el estado de cosas en que *a* es *F* y él puede utilizar varias distintas según cuál sea la actitud proposicional que tiene con relación a que *a* es *F*. Así, es posible imaginar un lenguaje en que representemos que *a* es *F* de una manera si *tememos* que *a* es *F*, de otra forma diferente si *esperamos* que *a* es *F* y de una tercera manera si *creemos* que *a* es *F*. Por ejemplo, podríamos imaginar lenguajes en que la *forma* de una oración incrustada en un verbo de complemento varíe según cuáles sean las actitudes proposicionales que exprese el verbo. Por lo que yo sé, no existen lenguajes de esta índole. Resulta sorprendente que no los haya.

Creo que esto abre líneas de especulación interesantes, pero no las voy a desarrollar aquí. Si la condición que acabamos de mencionar es razonablemente restringida, es lo suficientemente restringida para nuestros objetivos actuales.

Convendría insistir en que esta condición no tiene nada de trivial. Es fácil verlo si se piensa que *no* se cumpliría, por ejemplo, en el caso de alguien que utilizara ⌜*b* es *G*⌝ para representar el estado de cosas en que *a* es *F*. En este caso, sería «*b* es *G*» para representar el estado de cosas en que *a* es *F*. En este caso, sería ⌜*b* es *G*⌝ (y *no* ⌜*a* es *F*⌝) a lo que asiente si cree que *a* es *F*[12].

Lo que vengo a decir con todo esto es que alguien utiliza su lenguaje coherentemente cuando hay una cierta correspondencia entre lo que cree y la forma de las palabras que utiliza para expresar sus creencias. En el caso paradigmático —utilización de los términos en un lenguaje natural— esta correspondencia es válida porque el hablante sabe y respeta las convenciones que dirigen el lenguaje. Como veremos en el Capítulo 3, estas convenciones *son* fundamentalmente las reglas que emparejan actitudes proposicionales como las creencias con las formas de palabras que expresan esas actitudes. La clase de lenguaje privado que considera Wittgenstein se aparta de este paradigma en la medida en que la relación entre las formas lingüísticas y las actitudes proposicionales *no* está mediada por convenciones públicas. Por lo tanto, el desafío que el argumento del lenguaje privado plantea a la noción de un lenguaje del pensamiento es éste: Demuestra cómo es posible que dicha relación sea mediada por algo *que no sean* las convenciones públicas. Es lo que quiero hacer ahora con cierto detalle.

Todo mecanismo computacional es un sistema complejo que cambia su estado físico de alguna manera determinada por las leyes físicas. Es posible considerar este sistema como un ordenador en la misma medida en que es posible pensar en alguna reproducción cartográfica que empareje los estados físicos del mecanismo con las fórmulas de un lenguaje de computación, de tal manera que conserve entre las fórmulas las relaciones semánticas deseadas. Por ejemplo, podemos atribuir estados físicos de la máquina a las frases del lenguaje de tal forma que si $S_1 \ldots S_n$ son estados de la máquina, y si $F_1 \ldots F_{n-1}$, F_n son las frases emparejadas con $S_1 \ldots S_{n-1}$, S_n, respectivamente, la constitución física de la máquina sea tal que realizará esa secuencia de estados únicamente si $F_1 \ldots F_{n-1}$ constituye una prueba de F_n. Evidentemente, hay un número indefinido de maneras de emparejar los estados de la máquina con las fórmulas de un lenguaje que mantengan esta forma de relación, lo que equivale a decir que el desciframiento del código de la máquina tendrá una cierta indeterminación de traducción. Es también evidente que hay un número indefinido de maneras de atribuir fórmulas a estados de las máquinas que *no* mantengan tales relaciones entre las fórmulas, sólo que en estos casos no podemos interpretar como pruebas los cambios de estado de la máquinas.

Cuando concebimos un organismo como si fuera un ordenador, tratamos de atribuir a los estados físicos del organismo (por ejemplo, a los estados del sistema nervioso) fórmulas expuestas en el vocabulario de la teoría psicológica. En teoría, la ta-

---

[12] Estoy interpretando «cree» en sentido opaco en (*S* utiliza ⌜*a* es *F*⌝ para representar el estado de cosas en que *a* es *F*) si y sólo si (*S* asiente a las afirmaciones hechas utilizando la forma de las palabras ⌜*a* es *F*⌝ si y sólo si *S* cree que *a* es *F*). Lógicamente, esto da lugar a una interpretación igualmente opaca de «representar», que a mí me parece la natural. Sin embargo, si alguien cree que *S* utiliza ⌜*P*⌝ para representar el estado de cosas en que *b* es *G* se deduce de que *S* utiliza ⌜*P*⌝ para representar el estado de cosas en que *a* es *F*, y el estado de cosas en que *a* es *F* = el estado de cosas en que *b* es *G*, entonces debería interpretar transparentemente el «creer» de la primera fórmula.

rea debería realizarse de tal manera que (algunas, al menos) de las secuencias de estados que están causalmente implicadas en la producción de la conducta se puedan interpretar como computaciones que tienen descripciones adecuadas de la conducta, considera ésta como su «última línea»[13]. Lo que queremos decir es que, en el caso de los organismos como en el caso de los ordenadores reales, si damos con la forma correcta de atribuir fórmulas a los estados será posible interpretar la secuencia de hechos que *causa* el output como una *derivación* computacional del output. En resumen, los hechos orgánicos que aceptamos como implicados en la etiología de la conducta tendrán dos descripciones teóricamente pertinentes en el caso de que las cosas salgan bien: una descripción física en virtud de la cual queden incluidos dentro de las leyes causales y una descripción psicológica en virtud de la cual constituyan pasos computacionales desde el estímulo a la respuesta. Lo mismo ocurrirá, lógicamente, con las representaciones proximales del estímulo y respuesta[14,15].

---

[13] En el caso normal, una descripción de la conducta es «apropiada» en la medida en que lo es la (o una) descripción que el organismo trataba de que satisfaciera la conducta. No tendría ningún sentido, por ejemplo, emparejar los gestos articulatorios de los hablantes de inglés con oraciones del inglés de tal manera que la forma acústica «It's raining» (=está lloviendo) fuera asignada a la oración «someone is standing on my foot» (=alguien me está pisando el pie). Aunque sería posible determinar este emparejamiento —aunque podríamos adoptar un esquema para traducir las verbalizaciones de unos y otros de tal manera que, según tal esquema, lo que uno dice que hace el sonido «it's raining» es *que* alguien le está pisando el pie—, el aceptar esta atribución complicaría enormemente el papel de la teoría psicológica que trata de relacionar las verbalizaciones que producen las personas con las intenciones con que las producen. Al menos la suposición de que las personas que dicen «it's raining» están utilizando la frase «it's raining» para decir que está lloviendo nos posibilita una explicación sencilla y convincente del hecho de que muchas veces dichas personas lleven paraguas.

[14] Dennett (1969) es bastante tajante con este punto de vista:

> Es posible, quizá, que el cerebro haya llegado a adquirir métodos de almacenamiento y transmisión que impliquen hechos o estructuras sintácticamente analizables, de tal manera que, por ejemplo, ciertos modelos de moléculas o impulsos pudieran ser instancias de palabras-cerebrales, pero aun en el caso de que hubiera semejante «lenguaje» o «código»... debería haber también mecanismos para «leer» e «interpretar» este lenguaje. Sin estos mecanismos, el almacenamiento y transmisión de cosas semejantes a oraciones en el cerebro sería tan inútil como decir «giddyap» a un automóvil. Estos mecanismos de lectura deberían ser a su vez sistemas de tratamiento de la información, y ¿qué vamos a decir de *sus* estados y hechos internos? ¿Tienen partes analizables sintácticamente? El retroceso debe terminar en algún punto con sistemas que almacenen, transmitan y procesen la información en forma no sintáctica (p. 87).

Pero, en realidad, el retroceso no tiene ni que empezar. El argumento está fundamentalmente desorientado pues presupone una imagen del sistema nervioso como si enviara órdenes que deban ser «leídas» y traducidas a acciones (o, en cualquier caso, a contracciones musculares) por algún *otro* sistema que intervenga entre los nervios eferentes y los efectores. Pero esta imagen no forma parte de la teoría. Por el contrario, lo único que hace falta es que las propiedades *causales* de estos fenómenos físicos interpretadas como mensajes del código interno sean compatibles con las propiedades *lingüísticas* que la interpretación atribuye a esos hechos. Así, si los hechos del tipo físico *P* se deben interpretar como órdenes enviada al sistema efector *E*, en ese caso sería mejor que, *ceteris paribus*, las ocurrencias de los hechos-*P* sean causalmente suficientes para activar *E*. (*Ceteris paribus* quiere decir: salvo avería mecánica y salvo hechos interpretables como contraórdenes derogatorias dirigidas a *E*). Si se cumple esta condición, es difícil ver dónde se da la necesidad de un mecanismo «inteligente» para «leer» los hechos-*P*. Y, si no se cumple, es difícil ver qué sentido podría haber tenido el interpretar en un primer momento los hechos-*P* como órdenes para *E*.

[15] Una pregunta ya tópica que parece destinada a poner en aprietos a los psicólogos del procesamiento de la información es la siguiente: «Si se está dispuesto a atribuir las regularidades de la conducta de los organismos a reglas que éstos siguen inconscientemente, ¿por qué no se dice (por ejemplo) que los planetas «siguen» la ley de Kepler al realizar sus órbitas en torno al sol?». El objetivo, evidentemente, es hacer ver

Las observaciones realizadas hasta ahora tienen validez independientemente de toda presuposición concreta sobre el contenido de las teorías psicológicas. En realidad, son válidas para *cualquier* sistema físico en la medida en que sus cambios de estado se interpreten como computaciones. Pero el tema central de la discusión del Capítulo 1 era que toda teoría psicológica que tenga alguna posibilidad de ser verdadera tendrá que atribuir un papel especial a los estados computacionales de los organismos; es decir, la forma en que la información es almacenada, computada, aceptada, rechazada o procesada de alguna otra manera por el organismo explica sus estados cognitivos y, en especial, sus actitudes proposicionales. Es decir, el psicólogo presupone que ciertos procesos orgánicos satisfacen descripciones como «almacenar, aceptar, rechazar, computar, etc., *P*» y que el organismo aprende, percibe, decide, recuerda, cree, etc., todo lo que hace *porque* almacena, acepta, rechaza o computa todo lo que hace.

No quiero ocuparme ahora de la rectitud de esta suposiciones. Como vengo diciendo constantemente, nuestras opciones parecen ser o tolerarlas o prescindir completamente de teorías en la psicología cognitiva. Tampoco quiero ocuparme por extenso de *cuáles* son los procesos computacionales que se podrían atribuir adecuadamente a los organismos. Pero creo que hay algunas condiciones generalmente aceptadas (aunque no explícitamente) que afectan a estas atribuciones y que nos aproximan al núcleo de las suposiciones metodológicas de la actual psicología cognitiva.

Son tres: primera, que los estados computacionales imputables a los organismos se pueden explicar directamente como relaciones entre el organismo y las *fórmulas:* es decir, las fórmulas del código interno. Por eso, por ejemplo, en la medida en que se pueda decir (en sentido amplio) que el organismo almacena la información de que *P*, se debe poder decir *(en sentido estricto)* que el organismo está en determinada relación computacional con la fórmula *P* (por ejemplo, la relación de almacenar *P*). La segunda suposición es que la clase de relaciones básicas, teóricamente pertinentes, entre el organismo y las fórmulas del código interno (es decir, la clase de relaciones que pueden ser constitutivas de los estados y procesos computacionales del organismo) es bastante reducida; en concreto, que es pequeña comparada con la clase de relaciones teóricamente pertinentes entre organismo y proposiciones. Finalmente, y esto es lo importante, que por cada actitud proposicional del organismo (por ejemplo, temer, creer, querer, aprender, percibir, etc., que *P*) habrá una correspondiente relación computacional entre el organismo y alguna(s) fórmula(s) del código interno de tal

---

que el único caso *real* de seguimiento de una regla es el caso en que los organismos articulados siguen conscientemente una regla. Lo que hacen los demás organismos es (no *seguir* reglas sino) sencillamente actuar en conformidad con ellas.

Después de lo que hemos dicho en páginas anteriores debe quedar claro cuál es el modo de responder a esta objeción. Lo que diferencia lo que hacen los organismos de lo que hacen los planetas es que una *representación de las reglas que siguen constituye uno de los determinantes causales de su conducta.* Sin embargo, por lo que nosotros sabemos, en los planetas no ocurre esto: en ningún momento de la explicación causal de sus giros se hace mención de una estructura que codifique las leyes de Kepler y les haga girar. Los planetas *podrían* haber actuado de esa manera, pero los astrónomos nos aseguran que no lo hacen. Por tanto, el sistema solar no es un sistema computacional, pero usted y yo, por todo lo que sabemos en este momento, sí que podríamos hacerlo.

manera que *(el organismo tiene la actitud proposicional si y sólo si el organismo está en esa relación)* es nomológicamente necesario[16].

Esto es una forma larga, pero creo que útil, de decir que lo que se trata de hacer en la psicología cognitiva es explicar las actitudes proposicionales del organismo haciendo referencia a sus (hipotéticas) operaciones computacionales, y que la noción de operación *computacional* se toma aquí al pie de la letra, es decir, en cuanto operación definida por relación a *fórmulas* (internas). Así, por ejemplo, supongamos que recordar *P* es una de las relaciones que una teoría psicológica razonable podría reconocer entre un organismo y (la proposición) *P*. Supongamos, también, que almacenar *F* es una de las relaciones computacionales que una teoría psicológica razonable podría reconocer entre un organismo y la fórmula interna *F*. Entonces sería (en el mejor de los casos) una verdad contingente —precisamente la clase de verdad contingente que trata de formular la psicología cognitiva— que el organismo recuerda *P* si, y sólo sí, el organismo almacena *F*[17].

---

[16] Es claro que esta tercera condición no se puede cumplir tal como aparece formulada; y, aunque creo que se puede enmendar de varias maneras, no voy a intentar elegir una de ellas. El problema es que algunos términos de actitud proposicional son «relacionales» en el sentido de que se aplican al organismo (no solamente en virtud de su estado computacional, sino) en virtud de la forma de ser del mundo. Es decir, hay ciertas actitudes proposicionales de las que *no se pueden* presentar condiciones suficientes en términos de procesos de datos internos del tipo que hemos estado considerando. Pensemos, por ejemplo, en *saber que a es F*. Evidentemente, no hay ningún organismo que sepa que *a* es *F* a no ser que ocurra que *a* es *F*. O también, si *a* es *F* no está, en general, determinado por una determinación del estado computacional del organismo. De ahí se deduce que no puede haber ninguna relación computacional con una fórmula tal que (un organismo sabe que *a* es *F*) si y sólo si (están en esa relación con dicha fórmula). Podrían hacerse las mismas observaciones en relación con (pero no únicamente en relación con) las actitudes proposicionales designadas por otros verbos como «lamentar», «percibir», «recordar», etc.

Como ya he señalado antes, existen varias formas de aclarar el tema, pero ninguna de ellas me parece indiscutiblemente la mejor. Por ejemplo, podría estipularse que las actitudes proposicionales no relacionales y sólo ellas quedan cubiertas por la tercera condición, quedando así el problema de determinar qué actitudes proposicionales son las relacionales. También se podría «construir», por así decirlo, una actitud proposicional no relacional correspondiente a cada una de las relacionales «prescindiendo», en la atribución de las últimas, de las condiciones que imponen estados, hechos o procesos no psicológicos. Así, en una primera aproximación, «creer razonablemente» corresponde a «saber» en el sentido de que un organismo cree razonablemente que *a* es *F* si y sólo si el organismo cumple todas las condiciones para saber que *a* es *F* excepto la condición de factividad. Dentro de la misma línea, «parece que ve» corresponde a ver, «parece que oye» corresponde a oír, etc. Evidentemente, no se tiene la garantía de que el inglés contenga un nombre para cada una de las actitudes proposicionales no relacionales, pero supongo que no puede ponerse ninguna objeción a la utilización de neologismos para determinar el dominio de una ciencia. (En realidad, y prescindiendo ahora de la dificultad que nos ocupa, no se puede esperar más que una correspondencia aproximada entre el inventario de actitudes proposicionales que reconocemos pre-teóricamente y las que las teorías psicológicas llegan a comprobar. Las ciencias suelen determinar su objeto material según va progresando). Puede encontrarse un desarrollo más amplio en Fodor (1968).

[17] No constituye ninguna tautología ni ninguna forma de definición estipulativa del término técnico «almacenar» el que los organismos recuerden lo que, y sólo lo que, es almacenado por su sistema nervioso. De hecho, ni siquiera es *verdad* que los organismos recuerden lo que, y sólo lo que, almacenan sus sistemas nerviosos. Por un lado, gran parte de lo que se recuerda es reconstrucción de fragmentos almacenados (cf. Bartlett, 1961; Bransford y Franks, 1971) y, por el otro, gran parte de lo que se almacena muchas veces no se puede recordar porque no se puede recuperar (cf. la superioridad de la memoria de reconocimiento al recuerdo libre). Por eso, la correspondencia no es completa en ninguna de las dos direcciones: el almacenamiento, probablemente, es esencial al recuerdo, pero no es ni necesario ni suficiente. A fortiori el recordar no tiene carácter de «criterio» para el almacenamiento.

Quisiera hacer otra observación. Vengo diciendo que las teorías de la psicología cognitiva tratan de explicar las actitudes proposicionales de los organismos, y que tratan de hacerlo de una determinada manera: es decir, presentando, para cada actitud proposicional, condiciones nomológicamente necesarias y suficientes desde el punto de vista de las relaciones computacionales entre el organismo y las fórmulas del sistema de representación interna. Esto nos sugiere la siguiente descripción ontológica: hay, por así decirlo, *dos* cosas —la relación de los organismos con las proposiciones y la relación de los organismos con las fórmulas— y estas dos cosas están dispuestas de tal manera que la segunda es causalmente responsable de la primera (por ejemplo, el hecho de que el organismo esté en cierta relación con las fórmulas es causa de que el organismo esté en cierta relación con las proposiciones). Puedo imaginar que habrá quien rechace esta descripción basándose en motivos metafísicos, es decir, en considerar las proposiciones (o, en cualquier caso, las relaciones con las proposiciones) como la roca firme en que se apoya la psicología.

La observación que quiero hacer es que es *posible* rechazar esta descripción y admitir al mismo tiempo la forma de explicación psicológica que he estado proponiendo. En concreto, se podría considerar que las fórmulas explicativas básicas expresan (no relaciones causales entre las relaciones con las fórmulas y las relaciones con las proposiciones, sino) identidades de eventos contingentes. Es decir, podría pensarse que las teorías cognitivas encajan en un esquema de explicación que tendría, aproximadamente, esta forma: *tener la actitud R a la proposición P es contingentemente idéntico a estar en relación computacional C con la fórmula (o secuencia de fórmulas) F.* Una teoría cognitiva, en la medida en que fuera al mismo tiempo verdadera y general, explicaría probablemente la productividad de las actitudes proposicionales suponiendo un número infinito de ejemplos de sustitución de este esquema: una para cada una de las actitudes proposicionales que pueda mantener el organismo.

Hemos llegado a lo que me parece el corazón de las cuestiones específicamente metodológicas relacionadas con las teorías cognitivas. Si estamos dispuestos a atribuir actitudes proposicionales a un sistema, podemos dar sentido a la afirmación de que ese sistema utiliza un lenguaje, y podemos hacerlo independientemente de si el sistema es o no una persona y de si el uso del lenguaje está mediado o no por las convenciones, y de si el lenguaje utilizado funciona o no como medio de comunicación. Lo que hace falta (y parece que todo lo que hace falta) es que se dé la clase adecuada de correspondencia entre las actitudes del sistema con las proposiciones y las relaciones del mismo con las fórmulas del lenguaje. (Si *S recuerda que a es F si, y sólo si, S almacena ⌜a es F⌝* es nomológicamente necesario, en ese caso *S* utiliza ⌜*a* es *F*⌝ para representar que *a* es *F*: o lo hace, al menos, en aquellos de sus procesos cognitivos que son procesos de memoria.) Ya hemos señalado que, en el caso de los len-

---

Una teoría cognitiva trata de describir las formas en que las actitudes proposicionales de un organismo dependen de sus procesos de datos, donde «procesos de datos» son secuencias de operaciones con las fórmulas del lenguaje interno. Lo que quiero señalar ahora es que muchas veces esto no es fácil de hacer y no es algo que se pueda conseguir por definición estipulativa. En realidad, quizá no sea ni *posible* el conseguirlo. No tenemos ninguna garantía a priori de que todos los estados cognitivos de un organismo se *puedan* explicar por referencia al subconjunto especial que se compone de relaciones entre el organismo y las fórmulas de su sistema representacional interno. Lo único que sabemos a priori es que la psicología cognitiva actual presupone que esto es cierto.

guajes naturales, la correspondencia entre la relación del hablante con las fórmulas y las actitudes que tiene hacia las proposiciones está mediada por su adhesión a las convenciones que dirigen el lenguaje. En el caso del código interno, está determinado probablemente por la estructura innata del sistema nervioso. Pero, por lo que yo puedo decir, tal diferencia no afecta fundamentalmente a la explicación de la representación que hemos ofrecido. En ambos casos las fórmulas del sistema representan lo que representan porque la relación entre la utilización de las fórmulas y las actitudes proposicionales del organismo es tal cual es.

Nos encontramos así en situación de poder decir, con cierto detalle, qué viene a ser la analogía entre representación «privada» y «pública». Si ⌜a es F⌝ es una fórmula en un lenguaje público, entonces (*S* utiliza ⌜a es F⌝ para representar que *a* es *F*) sólo en el caso de que (*S* cree que *a* es *F* sólo en el caso de que *S* asiente a ⌜a es F⌝). Como lo que relaciona la creencia de *S* de que *a* es *F* con su asentimiento a ⌜a es F⌝ (lo que hace que la bicondicional incrustada sea verdadera) será, en el caso de los lenguajes públicos, la adhesión de *S* a las convenciones del lenguaje, podemos reemplazar dicha condición por la condición *C*.

(C) (*S* utiliza ⌜a es F⌝ para representar que *a* es *F*) sólo en el caso de que ([*S* cree que *a* es *F* sólo en el caso de que *S* asienta a ⌜a es F⌝] es convencional).

Pensemos ahora en el caso en que ⌜a es F⌝ es una fórmula del código interno. Entonces habrá una condición que sea válida para la fórmula y que difiere de *C* únicamente en que: (a) «asiente a» está reemplazado por una secuencia de una o más de las relaciones básicas a partir de las cuales se construyen las relaciones computacionales con las fórmulas internas, y (b) «es convencional» queda sustituido por «es nomológicamente necesario».

De esta manera, tenemos una respuesta que ofrecer a lo que yo califiqué de desafío básico del argumento del lenguaje privado a la noción de un sistema de representación interna: presentar una explicación de las relaciones de representación en el caso de las fórmulas de ese sistema. Sigue siendo una cuestión abierta el tema de si la representación interna, interpretada así, se parece a la representación del lenguaje natural lo suficiente como para que ambas puedan ser llamadas representación «en el mismo sentido». Pero no creo que haya que preocuparse demasiado por la respuesta a ese interrogante. Existe una analogía entre las dos clases de representación. Como los lenguajes públicos son convencionales y el lenguaje del pensamiento no lo es, no es probable que haya algo *más* que una analogía. Si nos fijamos en lo que hay de semejanza, nos inclinaremos a decir que el código interno es un lenguaje. Si nos fijamos en lo que falta a la semejanza para ser total, nos inclinaremos a pensar que el código interno es, en cierto sentido, un sistema representacional, pero que no es un lenguaje. Pero ni en uno ni en otro caso ello afectará a lo que yo considero que es el tema que está realmente en discusión: si las suposiciones metodológicas de la psicología computacional son coherentes. Nada de lo que hemos dicho hasta ahora nos hace pensar que no lo sean. En concreto, nada ha invalidado la afirmación de que el aprendizaje, incluyendo el aprendizaje de una primera lengua, implica esencialmente el uso de un sistema representacional interno *no* aprendido. Como no hemos encontrado razones para pensar que tal opinión esté equivocada y dado que es, como he señalado repetidamente, la única con que contamos, parece conveniente examinar las

implicaciones de la suposición de que dicho punto de vista es cierto. Esta es la tarea que vamos a emprender.

COMO DEBE SER EL LENGUAJE PRIVADO

He tratado de hacer frente a algunas de las objeciones filosóficas más importantes que se podrían presentar contra la idea de interpretar literalmente la opinión de que el aprender un (primer) lenguaje presupone formular y confirmar hipótesis sobre las propiedades semánticas de sus predicados. He considerado importante defender la coherencia conceptual de esta opinión porque, por una parte, parecía ser empíricamente plausible y, por otra, si la aceptamos nos veremos obligados a suponer que los organismos capaces de aprender un lenguaje deben tener antes acceso a algún sistema representacional en que se puedan expresar dichas propiedades. De aquí en adelante daré la cuestión por vista. Lo que quiero aclarar es que, habiendo llegado hasta aquí, tendremos que avanzar todavía un gran trecho.

Si decimos que una *definición de verdad* para el lenguaje natural L es cualquier teoría que asocie las condiciones de verdad con cada uno de los infinitos predicados de L, las suposiciones que hemos estado defendiendo se pueden resumir así: aprender L implica (al menos) aprender su definición de verdad. Ahora bien, una forma de formular una definición de verdad (no la única, pero, en mi opinión, las diferencias no afectan a los argumentos que vamos a considerar) es ésta: distinguimos entre un conjunto finito de predicados *elementales* de L, para cada uno de los cuales se *enumera* su determinación adecuada, y un conjunto infinito de predicados *complejos* cuyas condiciones de verdad asociadas están determinadas por algún procedimiento recursivo que especifica la definición de verdad. Los predicados distinguidos de esta manera suelen dar lugar a varias suposiciones. En primer lugar, todo predicado de L es o elemental o compuesto y ninguno es ambas cosas. En segundo lugar, todo predicado compuesto está formado por predicados elementales en forma que debe ser explicitada por la definición de verdad. En concreto, las condiciones de verdad asociadas con cualquier predicado complejo se fijan dada una especificación de su descripción estructural sintáctica y de los predicados elementales que contiene. Esto quiere decir que todo predicado de L es o elemental o eliminable en favor de predicados elementales por medio de una bicondicional definitoria. Así pues, una definición de verdad para un lenguaje natural contiene una lista de representaciones que determinan las extensiones de sus predicados elementales y un conjunto de reglas para definir sus predicados complejos en términos de sus predicados elementales.

Consideremos, pues, un predicado P en el vocabulario elemental de L. En primer lugar, una teoría de la verdad para L incluirá una afirmación de la forma de la fórmula (1) de tal manera que (a) la fórmula (1) sea verdadera y (b) «Gx» sea una fórmula en el vocabulario del metalenguaje en que se formula la definición de la verdad.

1) $\ulcorner Py \urcorner$ es verdad si y sólo si Gx

De donde se deduce fácilmente que G debe ser coextensivo con P, pues, si no lo fuera, la regla de verdad para P no sería ella misma verdadera. Ahora bien, el punto de

vista que hemos adoptado es aquel que dice que aprender *L* es (o, en cualquier caso, implica) aprender una definición de verdad para *L*. Supongamos que la fórmula (1) es parte de esa definición de verdad. Entonces, aprender *L* implica aprender la fórmula (1). En concreto, aprender *L* implica aprender que ⌜*Px*⌝ *es verdad si y sólo si x es G* es verdad en todos los casos de sustitución. Pero hay que tener en cuenta que aprender eso podría significar aprender *P* (aprender lo que significa *P*) sólo para un organismo que ya entienda *G*, porque, y este punto es crucial, en la fórmula (1) *G* se *utiliza,* no se menciona. Por eso, si aprender *P* es aprender una fórmula del tipo (1), un organismo sólo puede aprender *P* si es ya capaz de utilizar al menos un predicado que sea coextensivo con *P,* es decir, *G*.

Llegamos así a lo siguiente: Si aprender un lenguaje es cuestión, literalmente, de hacer y confirmar hipótesis sobre las condiciones de verdad asociadas con sus predicados, aprender un lenguaje presupone la capacidad de utilizar expresiones coextensivas con cada uno de los predicados elementales del lenguaje que se está aprendiendo. Pero, como hemos visto, las condiciones de verdad asociadas con *cualquier* predicado de *L* se pueden expresar en términos de las condiciones de verdad asociadas con los predicados elementales de *L*[18]. Parece que habría que sacar la conclusión de que sólo se puede aprender *L* si se sabe ya algún lenguaje lo suficientemente rico como para expresar la extensión de cualquier predicado de *L*. O, dicho tendenciosamente, sólo se pueden aprender cuáles son las propiedades semánticas de un término si se sabe ya un lenguaje que contenga un término que tenga las mismas propiedades semánticas.

Es ésta una consecuencia que horrorizará a quienes opinan que aprender un lenguaje es aprender la definición de verdad de que éste dispone; y hasta tal punto que vale la pena detenerse a averiguar cómo ha sido posible pasarlo tan completamente por alto. Creo que la respuesta es clara: mientras que la opinión de que las teorías semánticas son, o implican, definiciones de verdad tiene una larga tradición en la filosofía del lenguaje, hasta hace poco los filósofos no han llegado a pensar que en el aprendizaje de un lenguaje puede estar implicado el *aprender* una definición de verdad. Esta diferencia es definitiva. Es de gran importancia tener claro cómo cambia la situación conceptual cuando a las condiciones de una definición de verdad añadimos el requisito de que exprese lo que el hablante/oyente aprende cuando aprende a hablar.

Supongamos que tenemos un metalenguaje *M* en que se formulan las condiciones de verdad de las oraciones del lenguaje objeto *L*. En todos los casos, *menos* en el de la psicología, es práctico e inofensivo suponer que el vocabulario elemental de *L* está incluido en el vocabulario de *M*. Es útil porque nos garantiza que, por cada predicado elemental de *L,* habrá al menos un predicado coextensivo de *M;* es decir, ese mis-

[18] En realidad, es precisamente porque esto es verdad por lo que las definiciones de verdad son candidatos plausibles de lo-que-se-aprende-cuando-se-aprende-*L*. Las definiciones de verdad tratan de responder a la pregunta: «¿Cómo se puede entender la infinidad de predicados de *L* partiendo de la base de una representación finita de *L*?». La respuesta que dan es: realizando una reducción (finita) de cualquier predicado complejo a otro que sea coextensivo y esté compuesto de predicados elementales y expresiones del vocabulario lógico. Las observaciones análogas valen, *mutatis mutandis,* para las teorías semánticas intensionalistas, es decir, las teorías que afirman que la relación semántica crítica es (no la equivalencia sino) la vinculación mutua o sinonimia.

mo predicado. De esta manera nos ofrece una especie de forma normal de representar las extensiones de los predicados elementales de *L*. Dicho en términos aproximados, para todo predicado *P*, la representación canónica de aquellas frases de las que es predicado será «⌜*y* es *P*⌝ es verdad si y sólo si *x* es *P*», donde el mismo predicado, es decir *P*, se menciona en la parte izquierda de la fórmula y se utiliza en su parte derecha.

Es inofensivo incluir el vocabulario elemental de *L* en el vocabulario de *M* porque en la fórmula que acabamos de citar la presencia de *P* a mano derecha resulta perfectamente visible. Dado que estas fórmulas siguen siendo verdaderas en caso de sustituciones de cualquier predicado coextensivo con *P* (y, a fortiori, en caso de sustitución de cualquier predicado lógicamente equivalente o sinónimo) tenemos la garantía de que *cualquier* representación correcta de la extensión de *P* no pasará de ser materialmente equivalente a la representación que brinda la teoría de la verdad. En concreto, cualquiera que sea la representación de la extensión de *P* que aprendan de hecho los hablantes de *L*, tenemos la seguridad de que no pasará de ser materialmente equivalente a la fórmula citada.

Pero supongamos ahora que queremos incrustar una teoría de la verdad en una descripción de la psicología de hablantes/oyentes, de tal manera que la teoría deba implicar un número infinito de fórmulas (verdaderas) como *F:*

(F) Un hablante de *L* entiende *'P'* si y sólo si (ha aprendido que «⌜*y* es *P*⌝ es verdad si y sólo si *x* es *G*» es verdad para todos los casos de sustitución.

Lo que hay que observar es que la ocurrencia de *G* en *F* (a diferencia de la ocurrencia de *G* en la primera fórmula) *no* es transparente. «Y ha aprendido que *x* es *P*», siendo *P* y *Q* coextensivos, no implica «*y* ha aprendido que *x* es *Q*». En efecto, en ese caso una fórmula como *F* sólo será verdad si *G* es un predicado en el lenguaje que los hablantes de *L* utilizan de hecho para representar las extensiones de los predicados en *L*. Pero, evidentemente, *P* no puede ser tal predicado si *P* es un predicado de *L* pues, por definición, *L* es el lenguaje que hay que aprender. Salta a la vista que no se pueden utilizar los predicados que se están aprendiendo para aprender los predicados que se están usando.

En resumen, lo que es práctico e inofensivo en las definiciones de verdad *sin más* (hacer que el mismo predicado aparezca mencionado a mano izquierda de una regla de verdad y utilizado a mano derecha) es lo único que *no* debe ocurrir en aquellas representaciones de las condiciones de verdad que se supone también que representan lo que el hablante/oyente debe aprender sobre su lenguaje. Es decir, lo único que *G* *no* debe ser, en una fórmula como *F*, es *P*. Y ello porque *F* sólo puede ser verdad si *'G'* denota algún predicado en un lenguaje que ya conozca *S*. Y, por hipótesis, los *S*s que están aprendiendo *L* no saben ningún lenguaje en que ocurran los predicados de *L*.

Podemos ahora hacer un resumen de nuestra línea argumental. O es falso que aprender *L* es aprender su definición de verdad, o es falso que aprender una definición de verdad para *L* implica proyectar y confirmar hipótesis sobre las condiciones de verdad de los predicados de *L*, o nadie aprende *L* a no ser que sepa ya algún lenguaje distinto de *L pero lo suficientemente rico como para expresar las extensiones de los predicados de L*. Yo considero que, en el estado actual de desarrollo teórico

del lenguaje y el aprendizaje (y salvo las reservas mencionadas en la primera parte de este capítulo) sólo es admisible el tercer miembro de la disyunción. De ahí se deduce inmediatamente que no todos los lenguajes que uno sabe son lenguajes que se han aprendido, y que al menos uno de los lenguajes que uno sabe sin haberlo aprendido es tan potente como cualquier lenguaje que pueda llegar a aprender.

Admito que estas conclusiones pueden parecer realmente escandalosas. Me sentiría inclinado a considerarlas como una *reductio ad absurdum* de la teoría de que aprender un lenguaje es aprender las propiedades semánticas de sus predicados, con la excepción de que nunca se ha propuesto ninguna alternativa seria a dicha teoría[19]. En consonancia con la metodología general de este estudio, trataré de sobrellevar con buen ánimo lo que no sé cómo curar. En concreto, seguiré suponiendo que aprender un lenguaje natural es aprender las reglas que determinan las extensiones de sus predicados y seguiré tomando en serio las consecuencias que se puedan seguir de este punto de vista.

Por ejemplo algunas opiniones, por lo demás de aspecto totalmente razonable, sobre la relación entre hablar y pensar quedan eliminadas inmediatamente por la consideración de que el lenguaje interno debe ser lo suficientemente rico como para expresar la extensión de cualquier predicado del lenguaje natural que se pueda aprender. Así, muchos teóricos han considerado plausible la idea de que hay ciertos pensamientos que sería imposible tener a no ser por el hecho de que se ha aprendido un lenguaje. Estas opiniones son muy explícitas en las obras de Whorf (1956) y de sus seguidores, y parecen ser el objeto de algunos epigramas de Wittgenstein, como el de que un perro no *podría* pensar: quizá llueva mañana. Mi opinión es que, aunque esto es cierto en algún sentido, en otro sentido igualmente importante no lo es. Es lo que voy a intentar demostrar.

En primer lugar, puede dar la impresión de que no he sido demasiado justo con la opinión de que el lenguaje natural *es* el lenguaje del pensamiento. Conviene recordar que la principal objeción a esta opinión fue sencillamente que no puede ser verdad en relación con aquellos procesos computacionales que intervienen en la adquisición del mismo lenguaje natural. Pero, aunque podría admitirse que las computaciones *iniciales* implicadas en el aprendizaje de un primer lenguaje no pueden realizarse en el lenguaje que se está aprendiendo, sin embargo se podría decir que, una vez que se ha conseguido cierto dominio del lenguaje, el niño avanza extrapolando sus propios esfuerzos: El fragmento de lenguaje interiorizado anteriormente es utilizado para aprender lo que resta. Este proceso lleva con el tiempo a la construcción de un sistema representacional más desarrollado que el inicial, y este sistema más rico sirve para tener pensamientos que el niño no habría llegado a tener por otros procedimientos.

Es indudable que ocurre algo *parecido* a esto. En el caso extremo, consultamos un diccionario sobre una palabra que no entendemos, y el diccionario nos dice, en nuestro propio idioma, lo que significa la palabra. Eso *debe* contar, al menos, como

---

[19] Quizá esto constituiría un buen motivo para volver a insistir en que la diferencia entre las explicaciones intensionalista y extensionalista de la semántica *no* está implicada en la actual argumentación. Las teorías intensionalistas llevan precisamente a las mismas conclusiones que acabo de extraer, y lo hacen siguiendo el mismo camino.

utilizar una parte del propio lenguaje para aprender otra parte. Y si el adulto puede hacerlo con el procedimiento relativamente explícito de consultar un diccionario, ¿por qué no va a hacerlo el niño con el procedimiento relativamente implícito de consultar el conjunto de conocimientos que tienen los adultos? En concreto, ¿por qué no va a utilizar sus observaciones de cómo se aplica un término para confirmar hipótesis sobre la extensión de dicho término? Y ¿por qué no se van a formular estas hipótesis en un fragmento del mismo lenguaje que está aprendiendo el niño, es decir, en esa parte del lenguaje que se ha llegado ya a dominar?

Esto comienza a parecerse a un dilema. Por una parte, al aprender un lenguaje, a veces *es* útil utilizar el lenguaje que se está intentando aprender. Pero, por otra parte, la línea argumental que vengo siguiendo parece indicar que *no puede* serlo, pues he dicho varias veces que no se puede aprender *P* si no se aprende algo parecido a «$\ulcorner P_y \urcorner$ es verdad si y sólo si *Gx*», y que no se puede aprender eso si no se es capaz de utilizar G. Pero supongamos que G es un predicado (no del lenguaje interno sino) del mismo lenguaje que contiene *P*. En ese caso G debe ser algo que se ha aprendido y, *ex hypotehesi,* para aprender G hace falta haber aprendido (para uno u otro predicado) que G se aplica si y sólo si ese predicado se aplica. Lo importante es que este nuevo predicado debe ser parte del lenguaje interno o algo que se pueda hacer remontar hasta un predicado del lenguaje interno por iteraciones del presente argumento. En ninguno de los casos ningún predicado que pertenezca al mismo lenguaje que *P* desempeña un papel esencial para medir el aprendizaje de *P*.

El problema está, lógicamente, en que la bicondicional sea *transitiva.* Por eso, si puedo expresar la extensión de G en términos de, por ejemplo, H, y puedo expresar la extensión de *P* en términos de G, en ese caso puedo expresar la extensión de *P* en términos de H (es decir, «$\ulcorner y$ es $P \urcorner$ es verdad si y sólo si *Hx*»). Por eso, la introducción de *G* parece que no nos ha proporcionado ninguna ventaja. Parece que no hay ningún procedimiento por el que la parte de un lenguaje natural que sabemos pueda desempeñar un papel esencial para mediar el aprendizaje de la parte del lenguaje que no sabemos. Paradoja.

En realidad, se trata de dos paradojas íntimamente relacionadas. Queremos dejar un margen a la posibilidad de que haya *algún* sentido en que se pueda utilizar una parte de un lenguaje para aprender otras partes, y al mismo tiempo dejar margen a la posibilidad de que haya *algún* sentido en el que tener un lenguaje pueda permitir tener pensamientos que de otra forma no podrían darse. Sin embargo, pareee que las opiniones que hemos estado proponiendo no admiten ninguna de las dos posibilidades: en un lenguaje natural no es posible expresar nada que no se pueda expresar en el lenguaje del pensamiento. Si eso fuera posible, no podríamos aprender la fórmula del lenguaje natural que lo expresa[20].

Afortunadamente, ambas paradojas son falsas y por razones esencialmente idén-

---

[20] Sólo conozco una obra de psicología donde se haya presentado esta cuestión. Bryant (1974) observa: «el problema principal de la hipótesis de que los niños comienzan a entender y utilizar relaciones que les ayudan a resolver problemas porque aprenden los términos comparativos adecuados, como "mayor", es que deja sin respuesta la cuestión fundamental de cómo aprendieron el significado de estas palabras en primer lugar» (p. 27). Este argumento se generaliza a *cualquier* proposición de que el aprendizaje de una palabra es esencial para mediar el aprendizaje del concepto que expresa la palabra.

ticas. Empezando con el caso del aprendizaje, el argumento demuestra lo siguiente: supongamos que *F* es un fragmento de inglés y que un niño ha dominado *F* y sólo *F* en el momento *t*. Supongamos que *F'* es el resto del inglés. El niño puede utilizar el vocabulario y sintaxis de *F* para expresar las condiciones de verdad para los predicados de *F'* sólo en la medida en que las propiedades semánticas de los términos de *F'* sean ya expresables en *F*. Lo que el niño no puede hacer es utilizar el fragmento de lenguaje que conoce para incrementar la capacidad expresiva de los conceptos de que dispone. Pero puede ser capaz de utilizarlo para *otros* objetivos, y el hacerlo puede ser, desde el punto de vista de los simples hechos empíricos, esencial para el dominio de *F'*. La posibilidad más obvia es utilizar *F* con intenciones nemotécnicas.

Una afirmación muy socorrida en psicología es la de que los recursos nemotécnicos pueden ser esenciales para un sistema de memoria restringida al tener que enfrentarse con tareas de aprendizaje. Si, como parece razonable suponer, las expresiones del lenguaje natural relativamente sencillas son muchas veces coextensivas únicamente con fórmulas muy complejas del código interno, se ve fácilmente cómo es posible que el aprendizaje de una parte de un lenguaje natural constituya una precondición esencial para aprender el resto: los fragmentos aprendidos anteriormente pueden servir para abreviar las complicadas fórmulas internas, permitiendo así al niño reducir las exigencias implícitas en la proyección, confirmación y almacenamiento de hipótesis sobre las condiciones de verdad de los items aprendidos más tarde. Esto es una cosa bien conocida en la enseñanza del vocabulario de los sistemas formales. Los conceptos complejos *no* se introducen directamente en formas primitivas, sino más bien mediante una serie de definiciones interconectadas. El objetivo de esta práctica es poner límites a la complejidad de las fórmulas a que hay que enfrentarse en una fase determinada del proceso de aprendizaje[21].

Consideraciones fundamentalmente semejantes nos indican cómo podría ocurrir en último término que haya pensamientos que sólo puedan darse en quien habla un lenguaje. Es cierto que en relación con cada uno de los predicados del lenguaje natural debe ser posible expresar un predicado coextensivo del código interno. De ahí no se sigue que en relación con cada uno de los predicados del lenguaje natural *que se puedan concebir* haya un predicado *concebible* del código interno. No es ninguna novedad que los items individuales del vocabulario de un lenguaje natural pueden codificar conceptos de enorme sofisticación y complejidad. Si los terminos del lenguaje natural pueden llegar a incorporarse en el sistema computacional por algo parecido a un proceso de definición y abreviación, es perfectamente lógico pensar que el aprendizaje de un lenguaje natural puede incrementar la complejidad de los pensamientos que podemos tener. Para creer esto, sólo hace falta suponer que la complejidad de los pensamientos que se pueden tener está determinada *(inter alia)* por algún meca-

---

[21] Estoy dando por supuesto —como hacen muchos psicólogos— que los procesos cognitivos explotan al menos dos clases de almacenamiento: una «memoria permanente» que permite un acceso relativamente lento a una información de volumen esencialmente ilimitado y una «memoria de trabajo» que permite un acceso relativamente rápido y un número muy reducido de items. Es probable, en el caso del último sistema, que la capacidad de utilizar un determinado conjunto de informaciones dependa críticamente de la forma en que se codifica la información. Neisser (1967) estudia el tema en profundidad. Aquí nos limitaremos a señalar que una de las formas en que las partes de un lenguaje natural pueden mediar el posterior aprendizaje lingüístico es ofrecer el formato de dicha codificación.

nismo cuyas capacidades son sensibles a la forma en que se expresan los pensamientos. Como hemos notado anteriormente, es muy plausible considerar que los mecanismos de la memoria tienen esta propiedad.

Por eso, no me veo obligado a afirmar que un organismo articulado no tenga *ninguna* ventaja cognitiva sobre un organismo inarticulado. Ni hay ninguna necesidad de rechazar la afirmación whorfiana de que las clases de conceptos que se tienen pueden estar profundamente determinadas por el carácter del lenguaje natural que se habla. Lo mismo que es necesario distinguir los conceptos que se pueden expresar en el código interno de los conceptos que se pueden dar en un sistema de memoria restringida que computa con el código, también es necesario distinguir los conceptos que se *pueden* tener (*con excepción* de la memoria) y los que se llegan a emplear realmente. Esta última clase es sensible a las experiencias particulares del usuario del código, y no hay ninguna razón de principio por la que las experiencias que intervienen en el aprendizaje de un lenguaje natural no deban tener una influencia especialmente profunda para determinar cómo se explotan los recursos del lenguaje interior[22].

¿Qué es lo que se niega, entonces? Dicho aproximadamente, que se pueda aprender un lenguaje cuya fuerza expresiva sea mayor que la de un lenguaje que ya se sabe. Dicho con mayor precisión, que se pueda aprender un lenguaje cuyos predicados expresen extensiones no expresables por los de un sistema representacional con el que se contaba con anterioridad. Y todavía con mayor precisión, que se pueda aprender un lenguaje cuyos predicados expresen extensiones no expresables mediante predicados del sistema representacional *cuya utilización media el aprendizaje.*

Ahora bien, aunque todo esto es compatible con el hecho de que haya una ventaja computacional asociada al hecho de conocer un lenguaje natural, es *in*compatible con el hecho de que esta ventaja sea, por así decirlo, por principio. Si es cierto lo que estoy diciendo, todas estas ventajas computacionales —todos los efectos facilitadores del lenguaje sobre el pensamiento— deberán ser explicadas por referencia a parámetros de la «actuación», como la memoria, la fijación de la atención, etc. Dicho de

---

[22] No obstante, convendría insistir en que hay un desacuerdo radical entre los puntos de vista que vengo proponiendo y los que suscriben los relativistas lingüísticos. Para autores como Whorf, la estructura psicológica del recién nacido sería difusa e indeterminada. Por tanto, el hecho del desarrollo que las teorías psicológicas deben explicar es la aparición de los compromisos ontológicos aparentemente ordenados de los adultos a partir del caos sensorial que se supone es característico de la experiencia preverbal del niño. Este orden ha de venir de alguna parte, y parecería razonable señalar el inventario de categorías léxicas y gramaticales de la lengua que aprende el niño, al menos en el caso de que el teórico acepte la opinión de que las regularidades cognitivas han de ser reflejos de las regularidades *ambientales.* Según esta explicación, los sistemas cognitivos de los adultos deben diferir aproximadamente en la misma medida y de la misma forma que lo hacen las gramáticas y léxicos de sus lenguas y, por lo que se refiere a la teoría, las lenguas pueden diferir ilimitadamente.

Sin embargo, en el caso del código interno se invierten todos los presupuestos. Se supone que el niño (en realidad, el organismo infraverbal de cualquier especie) aporta al problema de organizar sus experiencias un sistema representacional estructurado en forma compleja y determinado endógenamente. Así pues, podrían preverse semejanzas en la organización cognitiva incluso en márgenes amplios de variación ambiental. En especial, el teórico no está obligado a descubrir análogos ambientales a los sesgos estructurales que aparecen en la ontología del adulto. Está, pues, preparado para no dejarse sorprender por la aparente intertraducibilidad de los lenguajes naturales, la existencia de universales lingüísticos, y las amplias analogías entre psicología humana e infrahumana. (Véase Fodor et al., 1974, donde se estudia el tema más ampliamente).

otra manera: si un ángel es un mecanismo con una memoria infinita y atención omnipresente —un mecanismo en que carece de sentido la distinción actuación/competencia—, en mi opinión no serviría para nada que los ángeles aprendieran latín; el sistema conceptual de que dispongan tras haberlo hecho no es más potente que aquel con el que comenzaron.

A estas alturas debería estar ya claro por qué el hecho de que podamos utilizar parte de un lenguaje natural para aprender otra parte (por ejemplo, recurriendo a diccionarios monolingües) no constituye ningún argumento en contra de la opinión de que nadie puede aprender un lenguaje más potente que algún lenguaje que ya sabe. No se puede utilizar la definición *D* para comprender la palabra W a no ser que (a) «*W* significa *D*» sea verdad y (b) se entienda *D*. Pero si se cumple (a), D y W deben ser al menos coextensivos, y por eso si (b) es verdad, quien aprenda W aprendiendo qué significa *D* debe entender ya al menos una fórmula coextensiva con W, es decir, aquella en que se expresa *D*. En pocas palabras, aprender una palabra puede ser aprender lo que se dice de ella en una definición de diccionario *únicamente cuando se trata de alguien que entiende la definición*. Por eso, el recurrir al diccionario no demuestra, después de todo, que podamos utilizar nuestro dominio de parte de un lenguaje natural para aprender expresiones que no podríamos haber dominado de otra manera. Todo lo que demuestra es lo que ya sabemos: una vez que se es capaz de expresar una extensión, se está en situación de aprender que *W* expresa esa extensión.

Estamos ya en condiciones de poder entender la importancia de todo esto. Para ello, sólo tenemos que considerar algunas implicaciones para ciertas áreas de la psicología, como la teoría del desarrollo cognitivo.

Existen, en primer lugar, muchas cosas que la mayoría de los adultos pueden hacer y que, en cambio, no pueden hacer la mayoría de los niños. Muchas de ellas implican destrezas cognitivas como la resolución de problemas de un nivel avanzado, el reconocimiento perceptivo de objetos complejos, y el hablar un lenguaje natural. Es razonable suponer que una psicología cognitiva adecuada debería postular procesos evolutivos cuyo funcionamiento hace de intermediario para la consecución de estas destrezas. Ahora bien, si mi interpretación es correcta, buena parte de la psicología del desarrollo cognitivo, especialmente tal como se ha visto influenciada por Vygotsky, Bruner y, por encima de todos, Piaget, se ha ocupado de defender tres hipótesis interrelacionadas referidas a estos procesos.

1. El desarrollo de las capacidades cognitivas del niño manifiesta una descomposición en *etapas* razonablemente ordenada.
2. Estas etapas, aunque en primera instancia se caracterizan por referencia a capacidades conductuales específicas que manifiesta el niño, son fundamentalmente expresiones de los tipos de conceptos de que dispone, correspondiendo los sistemas conceptuales más débiles a las primeras etapas.
3. El aprendizaje media la progresión evolutiva de una etapa a otra[23].

---

[23] El «aprendizaje no implica necesariamente *condicionamiento* o *asociación*. Estoy utilizando, más bien, la noción de aprendizaje de conceptos examinada en el Capítulo 1: una alteración en el sistema conceptual ocasionada por el entorno se considera como experiencia de aprendizaje de conceptos únicamente

Formulando estas afirmaciones en los términos que hemos estado utilizando en las páginas anteriores, podríamos decir que el punto de vista que está en discusión es que las capacidades intelectuales evolutivas del niño reflejan cambios en la competencia más que (meros) cambios en la actuación. El niño de más edad es capaz de hacer más clases de cosas que el niño de menos edad, y ello no, por ejemplo, porque tenga más memoria computacional con la que actuar, o porque su capacidad de atención sea mayor, o porque tenga un conocimiento más amplio de algunos hechos; la diferencia es más bien intrínseca a la capacidad expresiva de los sistemas conceptuales disponibles en las distintas etapas del desarrollo.

De todos los teóricos cognitivos quizá sea Piaget el que describe más explícitamente el desarrollo del niño en cuanto que implica la asimilación de una serie de «lógicas» de creciente capacidad representacional. Por mencionar un ejemplo tomado casi al azar, Piaget postula un nivel de desarrollo cognitivo intermedio entre el período «sensorio-motor» (en que se establece por primera vez la permanencia del objeto)[24] y el período «operacional concreto» (en que el niño manifiesta por primera vez la conservación de cantidades)[25]. En esta etapa intermedia,

las relaciones de orden, por ejemplo, que en el plano sensorio-motor estaban totalmente inmersas en el esquema sensorio-motor, ahora se asocian y dan lugar a una actividad específica de «clasificación» y «ordenamiento». De la misma manera, los esquemas de subordinación que originalmente estaban sólo implícitos, ahora se presentan totalmente separados y dan lugar a

---

si *lo que* se aprende (según su descripción teóricamente pertinente) está en relación de confirmación con los hechos que hacen que eso sea aprendido (según sus descripciones teóricamente pertinentes). Es decir, se trata de aprendizaje de conceptos únicamente si implica la proyección y comprobación de hipótesis.

[24] Piaget parece afirmar que la ontología del niño es inicialmente fenomenalista: el concepto de un mundo que está poblado de objetos que siguen existiendo aun cuando estén fuera del campo sensorial del perceptor es característico del niño *post*-sensoriomotor y (de alguna manera) aparece como consecuencia de la integración y coordinación de reflejos sensoriomotores innatamente determinados ante el impacto de estimulaciones del entorno. De hecho, incluso esta manera de formularlo parece que no refleja muy bien la medida en que Piaget supone que está sin estructurar el universo perceptual del niño, pues Piaget niega explícitamente que en la etapa sensoriomotora se haya llegado a establecer la distinción entre el perceptor y los objetos de su percepción. En la medida en que la ontología de esta etapa se parece a algo de lo que han tratado los filósofos, habría que decir que a lo que más se acerca es al monismo neutro. Para un desarrollo más amplio, véase Capítulo 1, *La construcción de la realidad en el niño.* Aquí nos limitaremos a señalar que las pruebas empíricas primarias que se citan para justificar la atribución a los niños de puntos de vista fenomenalistas es el hecho de que no buscan los objetos ocultos, por ejemplo, objetos que se han quitado de su campo visual mediante la interpolación de una pantalla opaca.

[25] En el experimento clásico sobre la conservación de la cantidad, se enseñan al niño dos depósitos idénticos (A y B) que, según él mismo admite, contienen la misma cantidad de líquido. Luego, se hace que el niño observe mientras el contenido de los depósitos (por ejemplo, el B) se va vaciando en una vasija estrecha y alta (C). Luego se le pregunta: «¿Cuál tiene más, C o A?». El niño que no ha llegado a adquirir la conservación se caracteriza por su inclinación a considerar que C tiene más que A (probablemente basándose en que el nivel de líquido de C es más alto que el nivel de líquido de A). Parece que la explicación fundamental de la no conservación es la ausencia, en el sistema conceptual del niño, de la inversión de relaciones. En concreto, no llega a darse cuenta de que los efectos de la operación de pasar el agua de B a C se podrían invertir mediante la operación paralela de pasar el agua de C a B. Algunos han observado, con cierta justicia, que esta explicación es una petición de principio (véase Wallach, 1969). Lo que nos interesa ahora es sencillamente que constituye un ejemplo relativamente claro de cómo trata Piaget de explicar una incapacidad cognitiva específica recurriendo a lagunas específicas en el poder expresivo de la lógica que se supone está utilizando el niño.

una actividad clasificatoria distinta, y el establecimiento de correspondencias pronto adquiere caracteres muy sistemáticos: uno/muchos; uno/uno/; copia con original, y así sucesivamente (Piaget, 1970, p. 64).

Lo que nos interesa ahora es el intento de Piaget de explicar el modelo de capacidades e incapacidades que se suponen características de esta etapa por referencia a las propiedades formales del sistema conceptual con que se piensa que está dotado el niño:

Al observar esta clase de conducta nos encontramos innegablemente con la llegada de la lógica, pero deberíamos tener en cuenta que esta lógica está limitada en dos aspectos esenciales: el ordenamiento o clasificación o establecimiento de correspondencias no presuponen la reversibilidad, por lo que no podemos hablar todavía de «operaciones» (pues hemos reservado el término para procedimientos que tienen un inverso), y debido a ello, no existen todavía principios de conservación cuantitativa... Por eso, podríamos considerar esta etapa del desarrollo intelectual como una etapa «semi-lógica», en el sentido totalmente literal de que carece de una mitad, a saber, las operaciones inversas (1970, pp. 64-65).

Es la obtención de una lógica en que se pueda expresar el inverso de una operación lo que se dice que explica las capacidades características de la etapa siguiente:

Entre las edades que van aproximadamente de los siete a los diez años el niño entra en una tercera etapa de desarrollo intelectual que implica el uso de operaciones... Ahora distribuye las cosas en series y comprende que al alinearlas, por ejemplo, en orden de tamaños cada vez mayores las está ordenando al mismo tiempo por orden de tamaño decreciente; la transitividad de las relaciones como mayor que, etc., que anteriormente pasaba inadvertida o sólo se consideraba como un hecho concreto, es ahora algo de lo que se tiene conciencia explícita... se han establecido los principios de conservación que antes estaban ausentes... (Piaget, 1970, pp. 65-66).

y así sucesivamente.

Ahora bien, todo esto podría ser cierto. Podría resultar que las clases de sistema representacional que utilizan los niños fueran, desde el punto de vista de los principios, más débiles que la clase de sistema que utilizan los adultos, y que se pudiera elaborar una explicación razonable de las etapas del desarrollo cognitivo refiriéndose a los aumentos de la capacidad expresiva de estos sistemas. Lo que creo que no se puede conseguir, sin embargo, es que el aprendizaje de conceptos suministre los mecanismos para las transiciones de una etapa a otra. Es decir, si el desarrollo cognitivo del niño es fundamentalmente el desarrollo de sistemas representacionales/conceptuales cada vez más potentes, el desarrollo cognitivo no puede ser consecuencia del aprendizaje de conceptos.

Las razones ya deben resultar conocidas, pues son esencialmente las que dan lugar a la conclusión de que no se puede aprender un lenguaje cuyos predicados expresen extensiones inexpresables en un lenguaje que no se tenga anteriormente; la diferencia entre aprender un predicado y aprender un concepto no tiene importancia en relación con el presente argumento.

Supongamos, por ejemplo, que somos un niño de la etapa uno que intenta aprender el concepto *C*. Lo menos que tenemos que hacer es aprender las condiciones por

las que algo es un caso de (cae dentro de) *C*. Por eso, es de suponer que tenemos que aprender algo de la forma (*x*) (*x* es *C* si y sólo si *x* es *F*) donde *F* es algún concepto que se aplica siempre que se aplica *C*. Sin embargo, es claro que una condición necesaria para ser capaces de aprender *eso* es que el sistema conceptual propio contenga *F*. Consideremos ahora el caso en que *C* es, por así decirlo, un concepto de la etapa *dos*. Si algo es un concepto de la etapa dos, de ahí se deduce que no es coextensivo con ningún concepto de la etapa *uno;* de lo contrario, la diferencia entre las etapas no sería una diferencia en el poder expresivo de los sistemas conceptuales que caracterizan a las etapas. Pero si el niño de la etapa uno no puede representar la extensión de *C* por medio de algún concepto del sistema con que cuenta, no puede representarlo en absoluto, pues, por definición, su sistema conceptual *es* precisamente la totalidad de mecanismos representacionales que puede utilizar para el procesamiento cognitivo. Y si no puede *representar* la extensión de *C,* no puede *aprender C* pues, por hipótesis, el aprendizaje de conceptos implica proyectar y confirmar las bicondicionales que determinan la extensión del concepto que se está aprendiendo. Por eso, o bien las condiciones de aplicación de un concepto de la etapa dos se *pueden* representar por medio de un concepto de la etapa uno, en cuyo caso no existe un sentido claro en que el sistema conceptual de la etapa dos sea más fuerte que el sistema conceptual de la etapa uno, o bien hay conceptos de la etapa dos cuya extensión *no se puede* representar en el vocabulario de la etapa uno, en cuyo caso no hay posibilidad de que los aprenda el niño de la etapa uno.

Es en el segundo cuerno de este dilema donde se ve cogido Piaget. Según su forma de ver las cosas, algunos conceptos, como la conservación de la cantidad, no pueden ser aprendidos por el niño «preoperacional» porque la caracterización de la extensión de los conceptos presupone operaciones algebraicas que no se pueden realizar con la lógica preoperacional. Pero si el niño no puede ni *representar* las condiciones en que se conservan las cantidades, ¿cómo va a ser posible que llegue a aprender que *son* ésas las condiciones en que se conservan las cantidades? No es de extrañar que Piaget se detenga tan poco en el análisis de los procesos de «equilibrio» de los que se supone que realizan las transiciones de una etapa a la siguiente. De hecho, la exposición de Piaget sobre el equilibrio es, por lo que yo puedo saber, *totalmente* descriptiva; no existe ninguna teoría de los procesos por los cuales se puedan conseguir los equilibrios.

Piaget afirma en apariencia que el desarrollo de la inteligencia implica establecer una serie de estados de equilibrio entre las demandas del niño sobre el entorno y las demandas del entorno sobre el niño: en concreto, entre el repertorio de esquemas de respuesta que el niño impone al mundo y los rasgos objetivos del mundo en que deben actuar los esquemas. La idea básica es que los esquemas del niño se perfeccionan y diferencian como respuesta a los procesos ambientales objetivos y cuanto más se perfeccionan y diferencian los esquemas de respuesta, más objetiva es la percepción del entorno implícita en las modalidades de adaptación del niño.

En sus comienzos, la asimilación es esencialmente la utilización del entorno externo por el sujeto en orden a nutrir sus esquemas hereditarios o adquiridos. No hace falta decir que esquemas como los de la succión, la vista, la prensión, etc., deben acomodarse constantemente a las cosas, y que las necesidades de esta acomodación frustran muchas veces los esfuerzos de asimila-

ción. Pero esta acomodación sigue tan indiferenciada de los procesos de asimilación que no da lugar a ninguna pauta de conducta activa especial, sino que consiste meramente en una adaptación de una pauta de conducta a los detalles de las cosas asimiladas... Por otra parte, en la proporción en que los esquemas se ven multiplicados y diferenciados por sus asimilaciones recíprocas así como por su progresiva acomodación a las diversidades de la realidad, la acomodación se disocia de la asimilación poco a poco y al mismo tiempo asegura una delimitación gradual del entorno externo y del sujeto... En exacta proporción con el progreso de la inteligencia en dirección a la diferenciación de los esquemas y a su asimilación recíproca, el universo avanza desde el egocentrismo integral e inconsciente de los comienzos a una solidificación creciente y a la objetivización (1954, pp. 351-352).

La orientación general de este tipo de explicación resultará conocida a los lectores de Dewey, para quien la función de la inteligencia es también conseguir una correspondencia cada vez más realista entre las acciones del organismo y los rasgos objetivos del mundo en que actúa.

Lo que queremos decir con todo esto es que, prescindiendo de la opinión que nos merezcan estos puntos de vista, lo que es evidente es que lo que falta en la versión piagetiana es una teoría que explique *cómo* consigue el organismo diferenciar sus esquemas *en la debida dirección,* es decir, en una dirección que, en general, *incremente* la correspondencia entre la imagen del entorno que implican los esquemas y las propiedades que tiene realmente el entorno. Si es cierto lo que he dicho más arriba, los puntos de vista de Piaget le *imposibilitan* para presentar esta teoría, pues, por una parte, quiere que la diferencia característica entre los niveles de equilibrio (es decir, entre las etapas del desarrollo) consista en la capacidad expresiva de la «lógica» que invocan, y, por otra, quiere que el mecanismo de equilibrio sea el aprendizaje. Como hemos visto, estos dos «desiderata» no se pueden satisfacer simultáneamente[26].

Hasta ahora he interpretado a Piaget en el sentido de que afirma que la diferencia subyacente entre las distintas etapas está en la capacidad expresiva de los sistemas conceptuales disponibles. Por lo tanto vale la pena señalar que el texto invita[27] a veces a una interpretación diferente. Según esta nueva perspectiva, la diferencia entre las etapas reside no en los conceptos que se pueden expresar, sino en la variedad de experiencias *a través de las cuales* se pueden emplear los conceptos. Generalmente se traza una línea de separación entre una etapa en la que los conceptos se aplican únicamente a lo que está realmente dentro del campo perceptivo y una etapa posterior en la que se extienden a los objetos que se imaginan sin ser percibidos. Es significativo el siguiente párrafo:

... la quinta etapa representa un progreso considerable con relación a la construcción del espacio; con la elaboración de grupos objetivos de desplazamientos que definen el comienzo de es-

---

[26] Dewey, por cierto, tiene una explicación explícita de los procesos mediante los cuales las creencias del niño convergen en una representación objetiva de su entorno: a saber, que son procesos de formación y confirmación de hipótesis. Esta postura es coherente en el caso de Dewey precisamente porque, a diferencia de Piaget, no admite el punto de vista de que las etapas evolutivas relativamente tempranas corresponden al empleo de una lógica relativamente empobrecida.

[27] Si es que es posible utilizar el término sin tratar de ironizar una forma de escribir como la de Piaget. La exégesis de Piaget no es clara. Espero que lo que he expuesto coincida con las intenciones de los textos, aunque no me sorprendería demasiado comprobar que no es así.

te período se puede decir, en efecto, que se establece el concepto de espacio experimental. Todo lo que es objeto de percepción directa (prescindiendo de los errores de hecho, claro está) puede ser organizado en un espacio común o en un entorno de desplazamiento homogéneo. Además, el sujeto toma conciencia de sus propios desplazamientos y de esta manera los localiza en relación con los demás. Pero su elaboración intelectual de las percepciones espaciales no trasciende todavía a la percepción misma para dar lugar a una verdadera representación de los desplazamientos. Por una parte, el niño no tiene en cuenta los desplazamientos que ocurren fuera del campo visual. Por la otra, el sujeto no se representa a sí mismo todos sus movimientos, fuera de la percepción directa que tenga de ellos (1954, p. 203).

Mi opinión personal, por si sirve de algo, es que Piaget postula realmente dos clases distintas de diferencias entre las etapas del desarrollo; dos aspectos en que los cambios de etapa pueden implicar un incremento de la capacidad expresiva del propio sistema conceptual. En uno de los casos, los cambios de etapa corresponden al empleo de sistemas conceptuales cada vez más poderosos dentro de un dominio determinado. En el otro, corresponden a la aplicación de un determinado sistema conceptual a la organización de fenómenos en dominios nuevos. Lo que quiero subrayar ahora, sin embargo, es que los mismos argumentos que demuestran que el aprendizaje no puede ser el mecanismo de la primera clase de transición de una etapa a otra demuestran, y con la misma fuerza, que tampoco puede ser el mecanismo de las transiciones de la segunda clase, en la medida en que supongamos que las transiciones de una etapa a la siguiente aumentan la capacidad expresiva del propio sistema conceptual. Es de suponer que el hecho de aprender que el concepto *C* se aplica en el dominio *D* es aprender que hay casos individuales en *D* que caen (o podrían caer) dentro de *C*. Pero, por hipótesis, aprender *eso* es cuestión de proyectar y confirmar una hipótesis, es decir, la hipótesis de que $(\exists x)$ (x está en D y (posiblemente o de hecho [Cx])). Sin embargo, a primera vista se aprecia que es imposible proyectar o confirmar dicha hipótesis a no ser que se pueda representar el estado de cosas en que algún individuo de *D* satisface *C*. Por eso, una vez más, el aprendizaje no incrementa la *capacidad expresiva* del propio sistema de conceptos (interpretados como conjunto de estados de cosas que uno puede representar) aunque, naturalmente, puede aumentar y muchas veces aumenta la propia información sobre qué estados de cosas predominan de hecho.

Me temo que habrá quienes piensen que todo esto es hablar por hablar, que estamos dando demasiadas vueltas sobre un tema tan *insignificante*. Por eso me voy a permitir proponer un ejemplo (no piagetiano) que aclare las complicaciones en que se ha metido Piaget.

Supongamos que tengo un mecanismo programado con las reglas de formación, axiomas y reglas de inferencia de la lógica proposicional clásica. Y supongamos que se me ocurre utilizar este mecanismo (de alguna manera) como modelo para el aprendizaje de la lógica de cuantificación de primer orden. (Elijo este ejemplo porque hay un sentido bien claro en el que la lógica de cuantificación de primer orden es más fuerte que la lógica proposicional: todo teorema de la primera es un teorema de la segunda pero no viceversa). ¿Cómo podría conseguirlo? Respuesta: no podría. Mi mecanismo no podrá aprender la lógica de la cuantificación a no ser que pueda aprender al menos las condiciones de verdad de fórmulas como *(x) Fx*. Pero mi pequeño modelo de aprendizaje no puede representarlas precisamente *porque* la lógica propo-

sicional es más débil que la lógica de cuantificación. Lo más que podría conseguir sería asociar *(x) Fx* con la conjunción indefinida *Fa & Fb & Fc...,* donde «...» equivaldría tácitamente a abandonar el proyecto.

Naturalmente, existen algunos procedimientos por los que mi mecanismo podría conseguir entender los cuantificadores y, entre éstos, hay algunos que tienen en común con el aprendizaje de conceptos el hecho de que las variables del entorno están implicadas de forma esencial. Por ejemplo, el dejarlo caer o golpearlo con un martillo podría producir la clase adecuada de cambios fortuitos en su estructura interna. Otra posibilidad es que los procesos físicos que intervienen en el mecanismo pudieran alterar ocasionalmente sus conexiones en la forma requerida sin la intervención de inputs del entorno. Pero lo que *no podría* ocurrir es que el mecanismo utilice el sistema conceptual disponible para *aprender* uno de más capacidad. Es decir, lo que no puede ocurrir es que pase de la fase uno a la fase dos mediante algo que pudiéramos reconocer como procedimiento *computacional.* En resumen, podría lograrlo un trauma, o la maduración, pero no el aprendizaje[28].

Existen muchas alternativas a la versión de Piaget que nos permiten preservar la suposición de que el desarrollo cognitivo se descompone en etapas. Por ejemplo, podría ser posible demostrar que el desarrollo cognitivo es, en último término, una cuestión de variables de la actuación más que de cambios en la competencia conceptual subyacente. Bryant y Trabasso han demostrado recientemente que el nivel de rendimiento del niño en algunas tareas características de Piaget cambia cuando se modifican las exigencias que las tareas imponen a la memoria[29]. Lo que no sabemos es cuántos de los descubrimientos de Piaget se pueden explicar de esta manera[30].

---

[28] Una forma menos tendenciosa de decirlo sería referirse a que la función de los inputs del entorno podría ser la de *disparar* la reorganización interna que sea necesaria para la transición de una fase a otra. La «impronta» (véase Thorpe, 1963) parece constituir un buen precedente de esta clase de interacción organismo-entorno, pues el papel del estímulo, en este caso, parece ser principalmente el de liberar patrones de conducta innatamente estructurados que de lo contrario no presentaría el organismo. Lo que nos interesa destacar ahora es que hay que distinguir claramente entre esta forma de explotación de los inputs del entorno y lo que ocurre en cualquier variedad de aprendizaje de conceptos, pues, como observamos en el Capítulo 1, un apecto decisivo de esto último es que el conocimiento que el organismo tiene de su entorno se explota para confirmar (o des-confirmar) generalizaciones sobre las extensiones de los conceptos. En efecto, los estímulos disparadores pueden tener una relación *arbitraria* con las estructuras que liberan, pero en el aprendizaje de conceptos los datos del entorno deben estar en relación de *confirmación* con las hipótesis que seleccionan.

[29] Bryant y Trabasso (1971). En concreto, demostraron que los niños «preoperacionales» pueden hacer frente a las inferencias que afectan a la transitividad de la longitud con tal que estén intensamente adiestrados en las premisas de la deducción antes de que se les pida que extraigan la conclusión. Esto permite pensar que el problema no es que el sistema conceptual del niño no pueda expresar la noción de transitividad, sino más bien que la memoria computacional de que dispone el niño preoperacional no es lo suficientemente grande como para retener las premisas de las que se deducen las conclusiones de los argumentos de transitividad. El niño *es* capaz de llegar a la conclusión correcta si antes se establecen las premisas en un sistema de memoria lo suficientemente amplio como para retenerlas, es decir, en la memoria «permanente».

[30] Un experimento útil de la psicología evolutiva consiste en tratar de imaginar modelos que presenten discontinuidades por etapas en la *conducta* como consecuencia del incremento de parámetros de «actuación» tales como la capacidad de la memoria computacional. Es evidente que existen muchos sistemas de esta naturaleza. Imaginemos, por ejemplo, un mecanismo de comprobación de teoremas para la lógica proposicional cuya única «regla» tiene *n* items de longitud. Imaginemos que el límite de la memoria de cál-

Queda abierta la posibilidad de que el desarrollo cognitivo del niño sea un desarrollo conceptual pero que el cambio de un sistema conceptual más débil a otro más fuerte se realice mediante variables de maduración, análogas a una alteración de la estructura física de un ordenador. (No hace falta negar que el entorno puedc proporcionar inputs que son esenciales —e incluso específicos— para iniciar o apoyar estas reorganizaciones madurativas determinadas endogenamente.) También se pueden presentar versiones mixtas. Algunos de los sistemas computacionales de que dispone el niño pueden estar limitados única o primariamente por variables de actuación mientras que otros pueden madurar. Algo semejante es lo que nos sugiere la consideración de que la capacidad computacional relativamente limitada que manifiesta el niño en las situaciones explícitas de resolución de problemas del tipo que considera Piaget no es impedimento para el ejercicio de mecanismos computacionales enormemente potentes en procesos tan especializados como la integración motora, aprendizaje lingüístico, orientación espacial, y reconocimiento facial. Un corte temporal en el desarrollo cognitivo del niño podría presentar haces de mecanismos computacionales cada uno de los cuales estaría en una etapa *diferente* del desarrollo y cada uno de los cuales plantearía sus propias exigencias sobre el tipo y cantidades de inputs ambientales que es capaz de explotar. Ninguna de estas teorías sobre las etapas queda descartada por los argumentos que hemos expuesto anteriormente. Lo que *sí* demuestran los argumentos es precisamente que si hay etapas y éstas están determinadas por la capacidad expresiva del sistema conceptual subyacente, el mecanismo del desarrollo cognitivo no puede ser, por lógica, el aprendizaje de conceptos.

Podemos terminar este capítulo exponiendo una paradoja. Lo que hemos demostrado es esto: si el mecanismo del aprendizaje de conceptos es la proyección y confirmación de hipótesis (y qué otra cosa *podría* ser), existe un sentido en el que no puede haber nada que se pueda considerar como aprendizaje de un nuevo concepto. Si la referencia a la comprobación de hipótesis es cierta, la hipótesis cuya aceptación es necesaria y suficiente para aprender *C* es que *C* es el concepto que satisface las condiciones de individuación de ∅ para uno u otro concepto ∅. Pero un concepto que satisfaga las condiciones que individualizan a ∅ *es* el concepto ∅. De ahí se deduce que ningún proceso que consista en confirmar dicha hipótesis podría ser el aprendizaje de un concepto *nuevo* (es decir, un concepto distinto de ∅)[31]. Lo que debe ocu-

---

culo de la máquina viene dado por el número de items de las fórmulas presentadas, y que va aumentando con el paso del tiempo, comenzando en un valor menor a $m+n$, donde m es la fórmula más breve a que se aplica la regla. (En realidad, estamos imaginando que la memoria computacional se va haciendo mayor según va «desarrollándose» el mecanismo.) El output de este mecanismo manifestará una discontinuidad conductual en forma de etapas en cuanto que habrá un valor de t tal que todas las pruebas que dé antes de t contengan solamente secuencias de series de longitud creciente mientras que, después de t, la longitud de las series puede aumentar *o* disminuir dentro de una prueba determinada. Lo que nos interesa de este mecanismo, por lo demás totalmente carente de interes, es que puede servir de advertencia contra la suposición de que las discontinuidades conductuales deben atribuirse invariablemente a la intervención de procesos subyacentes que no experimentan incremento.

[31] Esta forma de decirlo no es realmente distinta de las que he utilizado anteriormente, aunque pueda *parecerlo*. Lo único que he hecho ha sido formular el argumento de forma que explicite su neutralidad en la controversia intensionalista/extensionalista sobre la individuación de los conceptos.

Supongamos que alguien adopta la visión extensionalista de los conceptos y supongamos, como siempre, que identificamos aprender el concepto *C* con aprender que (x) Cx si y sólo si Fx; es decir, que ser *F*

rrir en la tarea de «aprendizaje de conceptos» descrita en el Capítulo 1, por ejemplo, no es que se interiorice un concepto nuevo, sino sencillamente que el sujeto aprende *cuál* de los varios conceptos localmente coextensivos tiene valor de criterio para la ocurrencia de la recompensa. Dicho brevemente, la tarea de aprendizaje de conceptos no se puede interpretar coherentemente como una tarea en que se aprendan conceptos y como, aparte del aprendizaje de memoria, el «aprendizaje de conceptos» es la única forma de aprendizaje para la que la psicología nos ofrece un modelo, parece justo decir que si existe ese proceso de aprendizaje de un nuevo concepto, nadie tiene la menor idea de cómo podría ser.

Si esto *es* una paradoja, es precisamente la misma con que venimos tropezando en todo momento: el único sentido coherente que podamos dar a los modelos de aprendizaje que existen actualmente es un sentido que presupone un nativismo muy extremo. Y quizá esto no sea tan malo como parece, ya que se pueden proponer varias consideraciones para mejorar las cosas.

1. Es posible que los conceptos complejos (como, por ejemplo, «aeroplano») se descompongan en conceptos más simples (como «aparato volador»). En el próximo capítulo comprobaremos que esta opinión está muy de moda en las teorías semánticas más actuales; en realidad, en una u otra versión, ha estado presente al menos desde Locke. Pero, prescindiendo de todo eso, es posible que sea cierta, y, si lo es, puede ser de utilidad. Concedamos que nadie puede aprender qué es un aeroplano a no ser que tenga ya los conceptos de que se compone dicho concepto junto con todas las operaciones combinatorias sobre los conceptos elementales que son necesarias para componer «aeroplano». Pero, aunque se nos exija ser nativistas en ese sentido, podemos reconocer perfectamente que sólo experiencias como, por ejemplo, la de estar en contacto con aeroplanos o tratar de inventar una forma de volar, etc., podrían dar lugar a la construcción del concepto complejo pertinente. Si, en resumidas cuentas, existen conceptos elementales por medio de los cuales se pueden especificar todos los demás, sólo hay que suponer que no se aprenden los primeros. En este sentido, el «aprendizaje de conceptos» se puede reconstruir como un proceso en que se componen conceptos complejos nuevos partiendo de sus elementos previamente dados. (Puede verse una exposición reveladora de este planteamiento de la «química mental» en la psicología del aprendizaje de conceptos en Savin, 1973).

2. La opinión que venimos proponiendo no exige que el sistema conceptual innato deba estar literalmente presente «al nacer»; sólo requiere que no sea aprendido.

---

es necesario y suficiente para ser *C*. Como *C* y *F* son conceptos coextensivos y como, según la hipótesis extensionalista, los conceptos coextensivos son idénticos, el concepto *C* = el concepto *F*. El mismo tipo de argumentación encajaría en una explicación intensionalista, con la diferencia de que el material bicondicionado tendría que ser fortalecido de forma aproximada para que dé lugar a un criterio en relación con el aprendizaje de *C*.

Hay que hacer notar que esta paradoja no se presenta con los *predicados;* aprender un predicado no es aprender qué predicado *es,* sino qué propiedades semánticas *expresa*. Dicho con menos misterio, si aprendo que el *predicado P* se aplica a *x* si y sólo si $\varnothing x$, aprendo una información totalmente contingente sobre la *forma lingüística «P»*. Los predicados se distinguen de los conceptos en que las condiciones para individualizar a los primeros hacen referencia a la sintaxis y vocabulario en que están formulados. Los predicados sinónimos son distintos aunque expresen el mismo concepto. Por consiguiente, los predicados distintos pueden tener propiedades semánticas idénticas. En cambio los conceptos distintos no las pueden tener.

Quizá sea un consuelo poco convincente, pero creo que los hechos no dan para más.

3. Es posible que el entorno desempeñe en la determinación del carácter del propio repertorio conceptual un papel muy distinto del que desempeña en la fijación del conjunto de conceptos que contiene el propio repertorio; por ejemplo, que proporcione *ejemplares* de los propios conceptos. Insisto en ello porque puede ocurrir que todo lo que haya que decir de algunos conceptos (por ejemplo, «rojo»)[32] es que son los conceptos de algo suficientemente semejante a ciertos ejemplares designados. Cuando decimos esto, estamos diciendo que aprender el concepto «rojo» es aprender algo parecido a «(x) x es rojo si y sólo si x es suficientemente semejante a E_i», donde E_i designa un ejemplar del color como puede ser una amapola, una puesta de sol o una nariz cuando hace frío. Evidentemente, los inputs del entorno podrían representar una contribución esencial para *esta* forma de aprendizaje de conceptos, ya que suministran el ejemplar. Lo que interesa ahora es que el proceso por el que alguien se familiariza con el ejemplar no es en cuanto tal un proceso de formación y comprobación de hipótesis; es más bien un proceso de abrir los ojos y mirar.

¿Sirve de mucho esta consideración? Desde luego, mitigará las suposiciones nativistas sobre los *conceptos* a costa de las suposiciones nativistas sobre la semejanza. (No se puede utilizar *C es el concepto de cosas suficientemente semejantes a* E_i para aprender *C,* a no ser que se esté ya en situación de emplear *es suficientemente semejante a* E_i). Esto podría suponer una ventaja real si la noción pertinente de semejanza resultara ser sencilla y general. Sin embargo, si las formas en que las cosas caen dentro de un concepto por ser semejantes a los ejemplares de ese concepto resultan ser prácticamente tan variadas con los mismos conceptos, el recurso a la semejanza no supondrá ninguna reducción considerable de las suposiciones nativistas de la teoría del desarrollo. Creo que se trata de una cuestión empírica sin resolver, pero no soy optimista al respecto: en primer lugar, porque los recursos a la semejanza para definir las dimensiones en que se transfiere el aprendizaje han tenido hasta ahora una historia bastante deprimente entre las teorías psicológicas de la generalización; en segundo lugar, porque parece un hecho indiscutible que las formas en que las cosas se parecen entre sí no se parecen mucho entre sí. ¿Qué hay de común en lo que tienen en común los repollos o los reyes?

He estado insinuando algunas formas en que sería posible tener la esperanza de quitar mordiente al hecho de que no se pueda aprender un sistema conceptual más rico que el sistema conceptual con que se empieza, donde al aprendizaje se interpreta como un proceso de formación y confirmación de hipótesis. Quizá sería mejor poner fin a esta discusión insistiendo en que *hay* un sentido en el que se puede decir que la formación y comprobación de hipótesis no puede constituir una fuente de conceptos nuevos, lo mismo que hay un sentido en el que no puede contribuir al aprendizaje de predicados, a no ser los que son coextensivos con los que contienen las mismas hipótesis. Esto es, por así decir, una limitación intrínseca del modelo y, en cuanto tal, impone severas limitaciones a aquellas teorías del aprendizaje lingüístico o del desarrollo conceptual con las que es compatible el modelo. Creo que lo único que se puede

[32] Pero también, quizá, «vaca» y otros conceptos de clase. Puede verse una explicación filosófica en Putnam (en trámites de publicación) y en Kripke (1972). El lector interesado en una perspectiva psicológica sobre la relación entre ejemplares, estereotipos y conceptos de clase puede consultar Heider (1971).

hacer al respecto es acostumbrarse a aceptar el hecho. Indicamos en el Capítulo 1 que las teorías cognitivas que se conocen actualmente presuponen un lenguaje interno en que se realicen los procesos computacionales que postulan. Debemos añadir que los mismos modelos implican que ese lenguaje es enormemente rico (es decir, que es capaz de expresar cualquier concepto que el organismo pueda aprender o tener) y que su capacidad representacional está, en todos los aspectos, determinada innatamente. Así sea.

Capítulo 3

LA ESTRUCTURA DEL CODIGO INTERNO: ALGUNAS PRUEBAS LINGÜISTICAS

Tengo que utilizar palabras cuando hablo contigo.

T. S. ELIOT

No intentes nunca presentar las condiciones necesarias y suficientes para nada.

PROFESOR L. LINSKY
(en una conversación)

Las principales conclusiones de lo que hemos examinado hasta ahora son éstas:

1. Los modelos existentes sobre los procesos cognitivos caracterizan a éstos como fundamentalmente computacionales y por lo tanto presuponen un sistema representacional en que se realizan las computaciones.
2. Este sistema representacional no puede ser un lenguaje natural, aunque:
3. Las propiedades semánticas de todo predicado de un lenguaje natural que se pueda aprender deben ser expresables en el sistema representacional.

Estas reflexiones —si son ciertas— sirven para establecer una especie de límite inferior a la capacidad expresiva que debemos suponer que tiene el lenguaje del pensamiento. Pero nos dicen muy poco sobre el carácter detallado del sistema, y son precisamente estos detalles lo que quieren descubrir los profesionales de la psicología. En este capítulo y en el siguiente, consideraré varias formas de testimonio empírico que pueden ayudar a resolver esta cuestión. No obstante, los objetivos son modestos. Quiero tratar de convencer al lector de que la hipótesis del lenguaje interno no es «metafísica», en el sentido peyorativo del término; que hay consideraciones basadas en los hechos que permiten constreñir las teorías del código interno. Por consiguiente, me daré por satisfecho si se acepta que los *tipos* de argumentación que voy a intentar son pertinentes para la confirmación de dichas teorías. Considero que para encontrar ejemplos de ello que sean válidos y que se pueda demostrar que lo son hace falta un esfuerzo persistente y en colaboración entre varias disciplinas diferentes. En este momento no podemos imaginarnos a qué resultados nos hará llegar esta empresa.

He insistido en que el lenguaje del pensamiento no puede ser un lenguaje natural. Sin embargo, algunos hechos acerca de este último nos proporcionan los mejores datos para hacer inferencias sobre el primero. En la primera sección de este capítulo intentaré decir algo sobre las razones por las que esto es así. En las secciones posterio-

res propondré algunos ejemplos de argumentos en que partiendo de hechos de los lenguajes naturales se llega a teorías sobre el código interno.

No descubrimos nada al decir que la publicación, en 1957, de *Estructuras sintáticas,* de Chomsky, provocó una serie de cambios fundamentales en la forma en que los científicos consideran los lenguajes naturales y los procesos psicológicos que intervienen en su utilización. Quizá sea por la gran velocidad con que han cambiado las cosas en el campo de la lingüística y psicolingüística por lo que se ha prestado relativamente poca atención a la cuestión de cómo se articulan los modelos de lenguaje con las teorías de la cognición. Sin embargo, es una cuestión que hay que considerar con toda seriedad si nos proponemos utilizar los datos del lenguaje natural para controlar estas teorías. Lo que viene a continuación es un intento de ver, desde el punto de vista del psicólogo, sobre qué versa la nueva lingüística.

Los «enfrentamientos de paradigma», como saben todos los que tienen constumbre de asistir a cocktails, son confrontaciones difusas de cosmovisiones distintas. No se refieren a cuestiones concretas ni se resuelven mediante experimentos decisivos. Sin embargo, muchas veces es posible y útil describir los supuestos fundamentales en que discrepan los paradigmas. Si, en el caso que nos ocupa, quisiéramos decir en una frase qué es lo que la mayoría de los psicolingüistas aceptaban antes de la revolución chomskiana y han dejado de aceptar desde entonces, sería indudablemente la suposición de que una teoría del lenguaje es esencialmente una teoría de la causalidad de las verbalizaciones.

Es de suponer que las elocuciones tengan causas y no se pretende poner en duda que una psicología suficientemente desarrollada pueda, al menos en principio, identificar las condiciones causalmente necesarias y/o suficientes para las elocuciones que producen las personas. Sin embargo, de ahí no se desprende que la foma correcta (o al menos una forma útil) de taxonomizar las formas de elocución de un lenguaje sea la de agrupar aquellas cuya producción dependa de los mismos (o semejantes) estímulos. Quizá la aportación más importante de Chomsky a la teoría psicolingüística haya sido el haber caído en la cuenta de que tal inferencia es una conclusión errónea.

Antes de Chomsky, muchos psicólogos anglo-americanos suponían que las elocuciones se refieren a los estímulos que provocan su producción y, por lo tanto, que una teoría que agrupe tipos lingüísticos cuyas instancias tengan causas semejantes agrupará, ipso facto, estructuras que presenten al menos una propiedad semánticamente interesante: la correferencialidad[1]. La polémica de Chomsky (1959) contra Skinner es fundamentalmente una demostración de que:

1. Las variables del entorno que actúan sobre el hablante son únicamente una de las determinantes de lo que dice; entre las demás figuran sus conveniencias, sus creencias no lingüísticas y su información sobre las convenciones de su lenguaje.
2. Como la conducta verbal es producto de variables que se interrelacionan en

[1] La correferencialidad no era la única propiedad lingüística que se suponía caraeterizable de acuerdo con las condiciones comunes de elicitación. Se aceptaba de forma generalizada, por ejemplo, que la noción de que dos palabras pertenecen a la misma clase sintáctica (nombre, verbo, artículo, etc.) se podía reconstruir partiendo de la suposición de que se daba cierta superposición entre algunos de los estímulos que las producían. Puede verse una exposición más amplia en el Capítulo 2 de Fodor *et al.* (1974).

forma compleja, no hay ninguna razón para suponer que una taxonomía de las estructuras lingüísticas que esté basada en sus condiciones de elicitación conservará alguna de sus propiedades teóricamente interesantes.

3. Es claro, a posteriori, que no conservará la correferencialidad. La presencia de la cosa a la que se hace referencia entre los estímulos que provocan una elocución no es condición ni necesaria ni suficiente de su producción. Y, lo que es peor, en la medida en que *hay* coincidencia entre los estímulos y los referentes (como cuando alguien dice «mi nariz» refiriéndose a su nariz) es casi seguro que ello no tiene ninguna significación teórica: los mecanismos de que depende el uso referencial del lenguaje no necesitan tal coincidencia. (Págame bastante y me comprometo aquí y ahora a referirme a lo que se te antoje, pasado o presente, real o imaginario. Y «bastante» no tendría que ser mucho.)

La crítica de Chomsky es, en mi opinión, tremendamente radical y está perfectamente bien basada. No es simplemente, como han sugerido ciertos comentarios, que Chomsky esté interesado en una cosa (la estructura del lenguaje) y los psicólogos en otra (las variables del entorno que intervienen en la causalidad de la conducta verbal). La que dice Chomsky es que, tal como están las cosas, no hay ninguna razón para creer que haya parte alguna de la psicología, *incluyendo* el análisis causal de la conducta verbal, que encuentre utilidad en una taxonomía de las formas lingüísticas en clases cuyos miembros tengan en común sus condiciones de elicitación. Si la conducta verbal *es* realmente resultado de una interacción, sería de *esperar* que tal taxonomía no tuviera ninguna utilidad; las elocuciones que provoca un estímulo determinado pueden ser arbitrariamente heterogéneas según cuál sea el estado psicológico del organismo sobre el que actúa el estímulo.

Insisto en ello porque me parece que en muchos casos no se ha entendido bien el sentido del ataque de Chomsky. Por ejemplo, Judith Greene (1972) escribe:

> ... como ha señalado MacCorquodale (1970) en una valiente defensa de Skinner, Chomsky deja al hablante competente sin nada que decir. En la medida en que el *qué* de una respuesta verbal no se reduce a un «¡ay!» skinneriano ante un alfilerazo, tiene perfecto sentido preguntar en qué condiciones estimulares el hablante hará uso de su conocimiento de reglas lingüísticas complejas para producir una elocución determinada. De lo contrario, cuando Chomsky dice que la conducta lingüística es indeterminada incluso probabilísticamente, ¿se refiere a que nunca es exacto decir que hay algunas elocuciones que tienen más probabilidades que otras en un contexto concreto? Si los planteamientos de Chomsky y de Skinner no llegan a interrelacionarse en discusiones significativas es porque Chomsky no ve ningún problema en ello mientras que Skinner considera que ya lo ha resuelto (pp. 192-193).

Pero el desacuerdo entre Chomsky y Skinner *no* está en si la conducta verbal está o no causada (*los dos* suponen que lo está, y los dos lo hacen, supongo yo, por razones principalmente metafísicas). También hay que decir que la teoría de Skinner no es reducible a la observación de que sería bueno saber algo sobre la contribución de las variables del entorno a la causación de la conducta verbal. Tampoco es cierto que Chomsky sea insensible a la existencia de un problema en torno a *la forma* en que es causada la conducta verbal. Por el contrario, lo que Skinner (1957) trató de demostrar es que aprender un lenguaje es aprender las condiciones estimulares de las res-

puestas discriminadas, y por lo tanto que una teoría de la conducta verbal debe tratar las verbalizaciones *como* respuestas (es decir, debe definir sus generalizaciones en relación con clases de tipos lingüísticos cuyas instancias se produzcan en condiciones ambientales semejantes o idénticas). Lo que Chomsky defendía era que aprender un lenguaje *no* es aprender conexiones E-R, y por lo tanto que no es probable que una taxonomía de formas verbales atendiendo a los estímulos que las provocan sirve para comprender aspecto alguno del uso del lenguaje.

El error básico de Greene, como el de Skinner, es sencillamente el dar por descontado que la pregunta «¿Cuáles son las causas determinantes de la conducta verbal?» y la pregunta «¿Cuáles son los estímulos que provocan la conducta verbal?» son intercambiables. Y no lo son. Es muy probable que en la causalidad de las elocuciones estén implicados *todos* los estados y mecanismos psicológicos fundamentales (memoria, atención, motivación, creencia, utilidad, etc.). Por consiguiente, de la premisa de que las verbalizaciones son *causadas* no se puede deducir la conclusión de que las verbalizaciones son *respuestas.*

Si Chomsky está en lo cierto en todo esto (y creo que no hay motivos serios para ponerlo en duda), no se debe identificar el aprendizaje de un lenguaje con el aprendizaje de las condiciones estimulares bajo las que se producen las instancias de sus tipos. Y, si esto es cierto, se deduce que «¿Qué estímulo hizo que el hablante *S* produjera la elocución *U*?» no es el tipo adecuado de pregunta para un psicólogo interesado en la explicación de la conducta verbal. Entonces, ¿cuál el tipo de pregunta *adecuado*? Si las teorías del lenguaje no tratan del control de las elocuciones mediante el estímulo, ¿de qué tratan?

Desde la aparición de *Estructuras sintáticas,* la propuesta ortodoxa ha sido que las teorías lingüísticas son descripciones de lo que los hablantes oyentes saben sobre la estructura de su lenguaje y que las teorías psicolingüísticas son descripciones de los procedimientos por los que dicha información se despliega en la producción y comprensión del habla. Por razones que he expuesto en otras obras (Garrett y Fodor, 1968; Fodor *et al.,* 1974), no puedo decir que me entusiasme demasiado esta forma de entender la relación entre lingüística y psicolingüística. Y estoy convencido de que tiene serias limitaciones heurísticas en cuanto forma de iluminar la repercusión que los hechos relacionados con el lenguaje tienen sobre los objetivos generales de la psicología cognitiva. En las páginas siguientes, voy a proponer una forma, un tanto excéntrica, de interpretar la lingüística y la psicolingüística que se desarrollaron a partir de *Estructuras sintácticas.* En concreto, lo que voy a sugerir es que la mejor forma de entender toda esa investigación es considerarla como una aportación al desarrollo de una teoría de la comunicación verbal.

La cuestión fundamental a la que trata de dar respuesta una teoría del lenguaje es ésta: ¿Cómo es posible que los hablantes y oyentes se comuniquen mediante la producción de formas ondulatorias acústicas? O dicho más exactamente: en determinadas condiciones[2] la producción por el hablante *S* de un objeto acústico *U,* que es una instancia de un tipo lingüístico que pertenece al lenguaje *L,* basta para comunicar un

---

[2] Por ejemplo, que la elocución sea audible, que el oyente esté atendiendo, y así sucesivamente. A partir de ahora daré por supuesto que se cumplen estas condiciones previas.

mensaje determinado entre *S* y cualquier otro hablante de *L* situado convenientemente. ¿Cómo se puede explicar este hecho?

Está muy claro, pienso yo, cuál debe ser la forma general de la respuesta a esta pregunta. La comunicación verbal es posible porque, cuando *U* es una instancia de un tipo lingüístico de un lenguaje que pueden entender ambos, la producción/percepción de *U* puede dar lugar a un cierto tipo de correspondencia entre los estados mentales del hablante y el oyente. El objetivo último de una teoría del lenguaje es decir qué clase de correspondencia es ésta y describir los procesos computacionales implicados en su consecución. Quizá sea conveniente explicar todo esto con más detención.

Yo parto de que la esencia de la comunicación en un lenguaje natural es más o menos ésta: los hablantes producen formas ondulatorias que tienen como objetivo acomodarse a ciertas descripciones. Cuando las cosas van bien —cuando el hablante dice lo que quiere decir y el oyente entiende lo que se ha dicho en la forma que el hablante quería que se entendiera— la forma ondulatoria se acomoda a la descripción a la que pretendía acomodarse y el oyente se da cuenta de que se acomoda a esa descripción y de que pretendía acomodarse. Y, por sentido común: la comunicación sólo consigue su objetivo cuando el oyente deduce las intenciones del hablante a partir del carácter de la elocución producida.

No trato de hacer un análisis en regla de «S_1 comunicó *C* a S_2 produciendo la elocución *U*». Lo que pretendo es simplemente destacar el papel esencial de las descripciones a las que el hablante intenta que se acomoden sus elocuciones, y del reconocimiento por parte del oyente de que se acomodan de hecho a tales descripciones, para conseguir la comunicación verbal[3]. Es más fácil comprender este punto si pensamos en la comunicación escrita dentro de un lenguaje natural.

Todo lo que escribo en inglés tiene una descripción verdadera en un metalenguaje cuya operación sintáctica fundamental es la concatenación y cuyo vocabulario se compone de las letras *a-z* (inclusive) y ciertos signos de puntuación (por ejemplo, «(», «)», «, », «.», « », « « » », cuyos nombres serían respectivamente parénte-

[3] En Grice (1957) se presenta un análisis del «significado del hablante» que se mantiene más o menos dentro de estas líneas. Gran parte de lo que tengo que decir sobre las teorías del lenguaje en la primera parte de este capítulo es un intento de sugerir la forma en que podrían quedar encajadas en teorías de la comunicación que están ciertamente dentro de la orientación general de Grice, aunque no lo sigan con todo detalle.

Quizá sea conveniente dejar claro que esta forma de explicación tiene una interpretación natural en cuanto teoría *causal* de la comunicación. Si, tal como he supuesto, la elocución de una forma ondulatoria puede dar lugar a una cierta correspondencia entre los estados mentales del hablante y el oyente, podemos suponer que esto se debe a que, en los casos relevantes, la elocución es causalmente suficiente para iniciar la secuencia de procesos psicológicos del oyente que en último término acaban convirtiéndose en un estado mental que corresponde al del hablante. (Los hablantes/oyentes son sistemas computacionales *personificados,* y cualquier secuencia de hechos que constituya la codificación/decodificación de una elocución tendrá, probablemente, una descripción verdadera en cuanto secuencia de causas y efectos). Por ello, podría decirse que una condición necesaria y suficiente para la comunicación entre hablante y oyente es que los estados mentales del uno estén en la forma adecuada de relación causal con los estados mentales del otro. De la misma manera, una condición necesaria y suficiente para la comunicación lingüística en *L* es que sus instancias desempeñen el tipo adecuado de papel en las cadenas causales que median las relaciones causales entre los estados mentales de los hablantes/oyentes de *L*. Y una *teoría* de la comunicación en *L* es verdadera si y sólo si dice qué clase de papel es el tipo adecuado de papel.

sis a la izquierda, paréntesis a la derecha, coma, punto, espacio entre palabras y comillas)[4]. Así, si escribo «the dog» (=el perro), lo que escribo tiene una descripción verdadera en este lenguaje: la letra *t*, seguida de la letra *h*, seguida de la letra *e*, seguida de un espacio de separación entre palabras, seguido de la letra *d*..., etc. Además, estas descripciones sirven para individualizar los tipos en el siguiente sentido: toda descripción de estas características especifica plenamente el tipo a que pertenece una determinada instancia ortográfica con tal que consideremos que los tipos son secuencias de letras. (Si consideramos que los tipos son secuencias de palabras, este tipo de descripción no sirve para individualizar, pues una inscripción ambigua como «the bank» (= el banco) sólo recibe una descripción ortográfica a pesar de que es una señal de dos tipos distintos)*.

Lo que quiero dejar claro es que, aunque lo que escribo cuando escribo «the dog» tiene una verdadera descripción ortográfica, lo que trato de comunicar cuando escribo «the dog» no la tiene. De hecho, hay un sentido en el que no puedo utilizar el lenguaje ortográfico para referirme a aquello a lo que intento referirme cuando escribo «the dog», pues los símbolos del lenguaje ortográfico denotan letras y signos de puntuación, pero a los que trato de referirme cuando escribo «the dog» no es ni a una letra ni a un signo de puntuación sino a algún perro determinado en atención al contexto.

Así, cuando escribo «the dog» utilizo una secuencia ortográfica para referirme a algo que no es el «designatum» de esta secuencia. (Lo mismo podríamos decir del inglés hablado, con la única diferencia de que entonces el metalenguaje es fonético en vez de ortográfico). Sin embargo, esto no es ningún misterio; de hecho, es algo que está a la vista. Aunque lo que escribo cuando escribo «the dog» tiene una descripción verdadera en cuanto secuencia de letras, tiene *también* una descripción verdadera en cuanto expresión referente (es decir, la expresión que se compone —solamente— de la palabra inglesa «the», seguida de un espacio de separación entre palabras, seguido de la palabra inglesa «dog»; es decir, la expresión «the dog») y lo que designan las instancias de ese tipo de expresión (cuando designan algo) son perros. Debido precisamente a que «the dog» tiene una descripción verdadera en cuanto tipo de expresión cuyas instancias se refieren a los perros, los hablantes de inglés que tratan de designar a los perros ejecutan muchas veces instancias de ese tipo.

Podemos sacar alguna conclusión de estas consideraciones. En primer lugar, si concebimos la comunicación verbal como un proceso en el que el hablante produce emisiones que van dirigidas a acomodarse a ciertas descripciones y el oyente recupera las descripciones a que trataban de acomodarse las elocuciones, podremos constreñir de forma relevante la caracterización de las descripciones pertinentes. Por ejemplo, la descripción que intento que recupere el lector cuando lee mi inscripción «the dog» *no* es, en primer lugar, la descripción ortográfica; es más bien esta descripción en cuanto «expresión que se refiere a un perro determinado por el contexto». Si no tu-

[4] En beneficio de la sencillez, y para evitar molestar al lector más de lo necesario, omito los recursos de inscripción, como el subrayado, que no están concatenados con otros símbolos del vocabulario ortográfico.

* La palabra «bank» tiene en inglés dos significados fundamentales: establecimiento bancario y orilla de un río (N. de T.).

viera presente esta descripción al escribir «the dog», y si el lector no reconociera que la inscripción que he escrito se acomoda a esa descripción, es que no he conseguido comunicar una referencia al perro escribiendo «the dog».

En segundo lugar, aunque la descripción a que intentaba acomodar mi inscripción «the dog» no es, en *primer* lugar, su descripción en cuanto secuencia ortográfica, debería acomodarse *de hecho* a tal descripción, y debería reconocerse que lo hace, para así servir como vehículo para comunicar una referencia al perro a los lectores *de inglés.* Puedo escribir «le chien» con la intención de producir con ello una instancia de un tipo utilizado para referirse a perros determinados atendiendo al contexto. Pero si mi lector no sabe francés, no podrá recuperar la descripción que yo pretendía partiendo de la forma de la inscripción que he producido, y por lo tanto no se conseguirán los objetivos de la comunicación. En pocas palabras, si trato de comunicarme *en inglés,* debería poner los medios para que lo que escriba se acomode no sólo a la descripción adecuada en cuanto expresión de referencia sino *también* a las descripciones adecuadas en cuanto secuencia de letras, palabras, etc. inglesas. En definitiva, es precisamente porque lo que he escrito se acomoda a estas descripciones por lo que puede servir como vehículo de comunicación entre hablantes de inglés (no analfabetos).

En resumen, una de las cosas que comparto con otros miembros de mi comunidad lingüística es el conocimiento de las descripciones a que debe acomodarse una forma escrita para que sirva para comunicar referencias al perro a las personas que pertenecen a esa comunidad. En concreto, sé cuáles son las inferencias que puede esperarse que hagan, al encontrar las instancias que produzco, las personas hablantes de inglés en cuanto hablantes de inglés; inferencias que irán de la forma de mis inscripciones al estado de mis intenciones. Cuando he escrito «the dog» y consigo comunicar una referencia a un perro determinado por el contexto entra en juego esta clase de conocimiento: produzco una inscripción a partir de la cual un hablante de inglés en cuanto hablante de inglés puede deducir una referencia intencionada a un perro y los hablantes de inglés en cuanto hablantes de inglés deducen la referencia pretendida a partir de las propiedades lingüísticas de la inscripción que produzco.

Lo que estoy diciendo (yendo ya por fin al fondo de la cuestión) es que una forma adecuada de considerar el lenguaje natural es hacerlo al estilo tradicional: es decir, en cuanto sistema de convenciones para la expresión de intenciones comunicativas. Podría considerarse que las convenciones lingüísticas son una especie de libro de cocina que nos dice, para cada C que se puede comunicar mediante una expresión del lenguaje, «si quieres comunicar C, produce una elocución (o inscripción) que se acomode a las descripciones D_1, D_2... D_n» donde las instancias D_s podrían ser representaciones sintácticas, morfológicas y fonológicas de la elocución. Para el oyente valdrían las indicaciones contrarias: conocer las convenciones de un lenguaje es al menos saber que una elocución que se acomoda a D_1, D_2... D_n también se acomoda por norma general a la descripción «producida con la intención de comunicar C»[5].

[5] «Por norma general» significa más o menos esto: suponiendo que el hablante está utilizando el lenguaje de acuerdo con las convenciones. Un hablante *puede* utilizar una forma verbal con la intención de comunicar algo distinto de lo que generalmente se comunica con ella. Pero, si lo hace, está corriendo un riesgo: no puede suponer que alguien que sepa el lenguaje sea capaz ipso facto de construir sus intenciones

Todo esto nos lleva a un determinado modelo de intercambios comunicativos entre hablantes y oyentes que a mí me parece no sólo natural, sino inevitable. Un hablante es, por encima de todo, alguien con algo que trata de comunicar. A falta de un término más adecuado, voy a llamar a lo que tiene en mente mensaje. Si quiere comunicarse utilizando un lenguaje, su problema es construir una forma ondulatoria que sea una instancia del (o de un) tipo utilizado normalmente para expresar dicho mensaje en dicho lenguaje. Cuando las cosas marchan bien, lo que emite o escribe *será* una instancia de este tipo; e, incluso cuando las cosas no salgan bien, lo que escribe o dice tendrá la *intención* de ser una instancia de este tipo. Es decir, estará orientado a acomodarse a la descripción «una instancia del tipo utilizado normalmente para expresar el mensaje *M* en el lenguaje *L*».

En el caso paradigmático, el hablante será capaz de hacer frente a su problema precisamente porque *es* hablante. Ser hablante de *L* es saber lo suficiente de *L* como para ser capaz de producir la forma lingüística que los hablantes de *L* utilizan por norma general para comunicar *M*. Naturalmente, esto es una construcción muy idealizada. Es posible que no haya forma de comunicar *M* en *L*, en cuyo caso el hablante quizá tenga que recurrir a otro lenguaje, o a formas no lingüísticas de comunicación, o a formas de palabras que sólo se aproximan a sus intenciones comunicativas. Puede haber también más de una forma de comunicar *M* en *L*. De hecho, parece razonable suponer que si hay una forma habrá un número indefinido, y el hablante tendrá que escoger entre ellas. Esto significa, en efecto, que las intenciones del hablante no quedan totalmente descritas cuando se dice que trata de comunicar *M* emitiendo una instancia verbal que pertenece a *L*. Lo que dice realmente refleja una gama de preferencias estilísticas que pueden imponer limitaciones de distintos grados de sutileza a la forma de las palabras que elige. Sin embargo, lo importante es que en los casos paradigmáticos:

1. El hablante produce una forma ondulatoria.
2. La forma ondulatoria que produce será un caso de una forma de palabras utilizada por norma general para comunicar *M* en *L*.
3. El hecho de que haya producido esa forma ondulatoria (y no otra) se podrá explicar, en una primera aproximación, por referencia a los detalles de *M* y a las convenciones de *L*[6].

El oyente tiene el mismo problema, aunque visto desde el ángulo contrario. Dada una forma ondulatoria, debe determinar el mensaje que el hablante ha intentado co-

---

comunicativas. De hecho, generalmente *no* lo suponemos; más bien, suponemos que el oyente conoce no sólo las convenciones lingüísticas sino también muchas cosas sobre lo que es probable que cualquier persona racional *quiera* decir. Esta es la razón clásica por la que es difícil construir procedimientos formales para el análisis de contenido o la traducción de los textos del lenguaje natural. Véase, por ejemplo, el examen de este punto realizado por Bar-Hillel (1970).

[6] El modelo a que me he estado refiriendo está idealizado en otro sentido, ya que presupone que la elección por el oyente de un mensaje que comunicar es literal y totalmente anterior a su elección de la forma lingüística en que va a formular la comunicación. En los casos del habla considerada esto resulta, lógicamente, muy poco verosímil; lo que equivale a decir que cualesquiera que sean los mecanismos que median la traducción de mensajes a formas ondulatorias deben estar controlados por la actuación de bucles retroactivos.

municar al producirlo. También aquí, en los casos paradigmáticos, lo que sabe de su lenguaje será adecuado para conseguir la determinación. Naturalmente, lo que sabe de su lenguaje puede decirle más de lo que el hablante trató de comunicar. Por ejemplo, en la medida en que la elocución del hablante es una elección de entre las opciones estilísticas con que cuenta *L* para expresar *M,* lo que dice el hablante comunicará no solamente *M* sino también sus preferencias estilísticas, al menos si el oyente tiene sensibilidad para captarlo. Existe, por lo tanto, una acepción más bien amplia de «comunicar» en que las palabras pueden comunicar más de lo que alguien tenía intención de comunicar: se utiliza el verbo en sentido de «revelar». De aquí en adelante trataré de prescindir de este uso, pues me parece claramente inadecuado en los casos paradigmáticos en que el hablante produce una forma de palabras utilizada por norma general para comunicar *M,* mediante lo cual intenta comunicar *M.*

Alguien puede revelar su *penchant* a los galicismos poniendo en cursiva palabras como «penchant». Pero no se puede, en ese sentido, *revelar* la convicción de que está a punto de llover diciendo: «Está a punto de llover». No tiene sentido hablar de una manera en que se integren estos dos tipos de casos, y no parece muy plausible la opinión de que los casos del último tipo se reducen a los primeros. En términos generales, la comunicación es una de esas actividades en que las intenciones del organismo al producir la conducta cuentan entre los determinantes *lógicos* de la clase de conducta que se produce. Si no se tiene esto en cuenta, se aceptará la noción etológica de comunicación que, de hecho, comprende todos y cada uno de los intercambios de información entre organismos, por muy inadvertidamente que se haga; noción tan inflacionaria que es incapaz de soportar ningún peso teórico.

Así que tenemos un modelo: un hablante reproduce mensajes en formas ondulatorias, y un oyente reproduce formas ondulatorias en mensajes. El carácter de cada reproducción está determinado, *inter alia,* por las convenciones del lenguaje que comparten el hablante y el oyente. La comunicación verbal es posible porque el hablante y el oyente saben cuáles son estas convenciones y cómo han de utilizarlas: lo que sabe el hablante le permite tomar el valor de *U* que codifique un valor determinado de *M,* y lo que sabe el oyente le permite captar el valor de *M* que está codificado en un determinado valor de *U.* El ejercicio de su conocimiento consigue, por tanto, una cierta correspondencia entre los estados mentales del hablante y el oyente: el hablante puede construir elocuciones que expresan los mensajes que tiene intención de expresar; el oyente puede interpretar las intenciones comunicativas del hablante. En resumen, el hablante tiene en mente un valor de *M* y el oyente puede saber cuál es ese valor de *M*[7].

---

[7] Podríamos decirlo de otra manera: el problema del oyente es decidir qué hipótesis sobre las intenciones del hablante explica mejor su (del hablante) conducta verbal. En circunstancias normales, la suposición de que el hablante está siguiendo las reglas de su lenguaje constituirá una solución general para los problemas de esta índole. Así, por ejemplo, la mejor explicación de la conducta verbal de alguien que dice «Está lloviendo» será normalmente que trata de comunicar la información de que está lloviendo; la mejor explicación de la conducta verbal de alguien que dice «Tengo sólo una nariz» será normalmente que trata de comunicar la información de que tiene únicamente una nariz, etc.

Estas observaciones tienen el propósito de conectar esta discusión con una tradición dentro de la filosofía de la mente según la cual las atribuciones de estados mentales a los demás deben analizarse, por lo ge-

Muchas veces se señala que los planteamientos lingüísticos contemporáneos son «mentalistas». Lo que se suele querer decir con ello es que se supone que los items del vocabulario teórico de la lingüística y psicolingüística designan estados y procesos no conductuales. En este sentido, cualquier psicólogo que no sea conductista es ipso facto mentalista, y nunca se me habría ocurrido poner en duda seriamente que la teorización útil sobre el lenguaje tuviera que ser mentalista en este sentido. Sin embargo, el planteamiento que vamos a adoptar es mentalista en un sentido más fuerte. No sólo se asegura que los procesos no conductuales median la relación de comunicación entre el hablante y su oyente, sino también que la comunicación *consiste* realmente *en* establecer un cierto tipo de correspondencia entre sus estados mentales. Por eso me resulta alentador saber que esto es lo que todo el mundo ha pensado siempre que es la comunicación. Nos hemos comunicado cuanto tú me has dicho lo que tienes en mente y yo he entendido lo que tú me has dicho.

He comenzado esta exposición diciendo que quería demostrar cómo era posible considerar la reciente investigación lingüística y psicolingüística como una aportación a la teoría de la comunicación: de hecho, se trataba de ilustrar los objetivos de esta investigación encajándola en dicha teoría. Me parece importante hacerlo porque la teoría de la comunicación puede encajarse, de forma muy natural, en el tipo de descripción de los procesos cognitivos expuestos en los Capítulos 1 y 2. En la medida en que esta estrategia resulte positiva, deberíamos ser capaces de aclarar en gran medida el tema central de *este* capítulo: la repercusión de los hechos relacionados con los lenguajes naturales y el procesamiento del lenguaje natural en las teorías sobre el carácter del código central. La idea general es que los hechos relacionados con el lenguaje natural limitarán nuestras teorías de la comunicación, y a su vez las teorías de la comunicación limitarán nuestras teorías sobre las representaciones internas. Lo que me propongo ahora es avanzar un poco en esa dirección. En concreto, quiero hacer ver que hay una variedad de clases diferentes de condiciones que deberían ser satisfechas por una teoría convincente de los *mensajes,* y que esto es algo que tiene pleno sentido dado que la forma más plausible de interpretar los mensajes es considerarlos como fórmulas en el lenguaje del pensamiento.

Lo primero que hay que tener en cuenta es que lo que hayamos tenido que decir hasta ahora sobre la naturaleza de la comunicación verbal no implica ninguna opinión concreta sobre la estructura del lenguaje como no sea la verdad evidente de que, como las instancias lingüísticas son objetos acústicos, la comunicación verbal debe implicar la producción e interpretación de tales objetos. Lo que conecta la descripción de la comunicación que acabamos de presentar con la investigación actual sobre la estructura de los lenguajes naturales es la afirmación de que una gramática generativa de *L* especifica (algunas o todas) las descripciones a que se debe acomodar una instancia para conformarse a las convenciones lingüísticas de *L*. Dicho con palabras un poco distintas, dicha gramática especifica, para cada *M,* las descripciones (morfológicas, fonológicas, sintáticas, etc.) que debe satisfacer una instancia para pertene-

---

neral, como inferencias sobre la mejor explicación de su conducta: las atribuciones de intenciones comunicativas constituyen un caso especial en el que las conductas a explicar son (por ejemplo) verbalizaciones. Para una exposición de los problemas principales véase Putnam (1960b), Chihara y Fodor (1965), y Fodor (1968).

cer al tipo de oración que expresa *M* en *L*[8]. Como, según el modelo de comunicación que acabamos de proponer, una elocución servirá normalmente para comunicar *M* en *L* solamente si el oyente se asegura de (y el oyente reconoce) que la elocución se acomode a tales descripciones, podemos describir la conexión entre la teoría de la comunicación y la teoría de la gramática generativa haciendo referencia a dos hipótesis concretas:

1. La reproducción de mensajes en formas ondulatorias y viceversa es indirecta: las formas ondulatorias se emparejan con mensajes por medio de la computación de un número de representaciones intermedias.
2. Entre estas representaciones intermedias hay varias que corresponden a las descripciones estructurales de las oraciones que presentan las gramáticas generativas.

Consideradas simultáneamente, las hipótesis (1) y (2) equivalen a la afirmación de que las descripciones estructurales lingüísticas son «psicológicamente reales» y que «median» en el proceso de comunicación.

No voy a detenerme ahora a examinar las pruebas en favor de esta afirmación (véase una amplia exposición en Fodor *et al.,* 1974). Lo que quiero hacer es explicar la noción de descripción estructural hasta el punto que quede claro qué es lo que afirma dicha afirmación.

Toda gramática generativa de un lenguaje natural reconoce un conjunto fijo, finito de *niveles de descripción* en que reciben análisis las frases del lenguaje. Tradicionalmente (es decir, en los tipos de gramáticas inspirados por Chomsky, 1957) se proponen al menos los siguientes niveles: fonético, morfofonológico, sintáctico superficial y sintáctico profundo. Ahora bien, un determinado nivel de descripción puede ir asociado a un lenguaje formal. Es decir, cada nivel de descripción se puede identificar con un determinado (generalmente infinito) conjunto de fórmulas cuyos elementos están tomados del vocabulario del nivel y cuya sintaxis está determinada por las reglas de corrección del nivel. La población del nivel fonético, por ejemplo, consta de un conjunto infinito de secuencias de fonos concatenados. De forma análoga, la población del nivel sintáctico superficial está formada por un conjunto infinito de diagramas sintagmáticos de una sola raíz, cada uno de los cuales contiene un número finito de nodos de ramificación con nombres tomados de un vocabulario propio que incluye «sintagma nominal», «nombre», «verbo», «adjetivo», etc. Lo mismo ocurre, *mutatis mutandis,* con cada uno de los demás niveles de descripción.

Es condición de adecuación de una gramática generativa que cada oración del lenguaje que describe reciba al menos una representación (y como máximo un número finito de representaciones) en cada uno de los niveles de descripción que reconoce la gramática. Es decir, cada oración debe estar asociada a un conjunto de representaciones de tal manera que cada fórmula del conjunto esté bien formada en un nivel de descripción y que cada nivel de descripción aporte al menos una fórmula al conjun-

---

[8] En algunas ocasiones Chomsky ha dicho que una gramática de *L* especifica la correspondencia entre «forma y significado» para las oraciones de *L*. Véase especialmente la discusión del tema en Chomsky (1965).

to. Este conjunto de fórmulas es la *descripción estructural* de la oración relativa a la gramática[9].

He señalado antes que hay razones suficientes para aceptar una u otra versión de las hipótesis (1) y (2). Es decir: las computaciones subyacentes a la comunicación verbal especifican formas ondulatorias que corresponden a mensajes dados y mensajes que corresponden a formas ondulatorias dadas. Y, en el curso de este procesamiento, se computan una serie de representaciones intermedias, de las cuales hay al menos algunas que se corresponden estrechamente con las que recibcn las oraciones en los distintos niveles de descripción que reconocen las gramáticas generativas. Si esto es cierto, es ya una orientación sobre la forma en que los hechos relacionados con la estructura lingüística y los procesos lingüísticos pueden constreñir las teorías que tratan de especificar el carácter de los mensajes.

1. Nada puede ser representación adecuada del mensaje a no ser que pueda servir como input para un mecanismo que sea capaz de computar la descripción estructural de las oraciones que expresan ese mensaje; «descripción estructural» se entiende aquí en su sentido lingüístico técnico.
2. Nada puede ser representación adecuada de un mensaje a no ser que pueda ser producida como output por un mecanismo cuyo input es la descripción estructural de una oración que expresa el mensaje.

Lo que interesa retener, naturalmente, es que sabemos mucho sobre la forma de las descripciones estructurales y la información que contienen, y sabemos algo —aunque no mucho— sobre las clases de procesamiento de la información que se desarrolla en la codificación y decodificación de los objetos acústicos a los que se aplican las descripciones estructurales. Este tipo de información se refiere a la naturaleza de los mensajes pues, independientemente de lo que puedan ser los mensajes, éstos deben presentar una relación sistemática con las descripciones estructurales y esa relación debe ser computable mediante los procedimientos de manipulación de la información de que disponen los hablante/oyentes.

Pero, de hecho, podemos defendernos mejor. Hemos afirmado que las teorías sobre los mensajes se ven obligadas a ofrecer input/outputs para los modelos del hablante y el oyente, y que se trata de una construcción fundamental en la medida en que la investigación lingüística y psicolingüística sea capaz de proporcionarnos estos modelos. Pero si, como hemos supuesto, los mensajes especifican la información comunicada en los intercambios verbales, una descripción de la estructura de los mensajes debe satisfacer simultáneamente una serie de condiciones diferentes. Dicho con la mayor generalidad de que soy capaz, las estructuras que identificamos con los mensajes deberán ofrecer ámbitos adecuados para *cualquier* operación cognitiva que

---

[9] Dado que, por lo que podemos saber, los niveles lingüísticos son universales (es decir, dado que toda gramática empíricamente adecuada debe reconocer el mismo conjunto de niveles que cualquier otra) el afirmar que toda oración de todo lenguaje tiene una descripción estructural equivale a afirmar que toda oración de todo lenguaje tiene un deletreo fonético, un análisis morfofonológico, una estructura superficial, una estructura profunda, etc. Si, tal como yo he dado por supuesto, las descripciones estructurales son psicológicamente reales y median la relación de comunicación, la universalidad de los niveles descriptivos implica una universalidad correspondiente de los procesos psicológicos implicados en la producción y percepción del habla.

se aplique a la información que comunican las verbalizaciones. Una de estas operaciones es la codificación/decodificación de formas ondulatorias. Esta es precisamente la operación de que se ocupan fundamentalmente la lingüística y la psicolingüística, pero es evidente que no es la única.

Por citar un ejemplo elegido casi al azar, una de las cosas que somos capaces de hacer con la información transmitida lingüísticamente es compararla con la información que nos llega por medios no lingüísticos. Las cosas que los hablantes dicen se ven a menudo confirmadas, o refutadas, por las cosas que ven, oyen, saborean, tocan y huelen, y, probablemente, parte del conocimiento de un lenguaje consiste en saber que esto es así. En resumen, debe haber procedimientos computacionales que nos permitan utilizar lo que vemos desde la ventana para confirmar la observación de que está lloviendo, y estos procedimientos consiguen de alguna manera aplicarse simultáneamente a la información transmitida lingüística y visualmente. Una forma evidente de conseguir esto sería traducir todos los inputs perceptuales a un código común y luego definir la relación de confirmación para las fórmulas de ese código: tendríamos así un análogo exacto con lo que se intenta conseguir en la formalización de la relación de confirmación para las teorías científicas[10]. Es compatible con esta pro-

---

[10] Es obvio que existe una íntima relación entre teorías psicológicas de la fijación de la creencia y teorías filosóficas de la confirmación científica, y no sólo por la razón expresada en el Capítulo 2, de que ambas se ocupan del análisis de inferencias no demostrativas. Así, en psicología, pensamos que la disposición del sujeto a creer una afirmación está determinada, *inter alia,* por sus perceptos en curso. Y, en filosofía, consideramos que el grado de confirmación de una teoría científica está determinado, *inter alia,* por el carácter de los hechos que caen en el dominio de dicha teoría. Lo importante es que, en cada caso, tenemos una relación de confirmación que es válida, a primera vista, entre objetos «lingüísticos» (como afirmaciones y teorías) y objetos «no lingüísticos» (como lecturas de indicadores y perceptos). Esto plantea un problema en la medida en que una teoría de la inferencia científica, o de la fijación de las creencias, busca tratar la confirmación como una relación *formal* ya que, dicho suavemente, cuesta trabajo pensar en una idea de forma que hiciera formalmente comparables a los objetos lingüísticos y no lingüísticos.

La forma tradicional de afrontar este problema dentro de la filosofía de la ciencia ha consistido sencillamente en *presuponer* que tanto las hipótesis como los hechos que las confirman tienen representaciones canónicas en un lenguaje propio; por eso, la confirmación se define como una relación formal entre fórmulas de ese lenguaje si es que resulta posible de alguna manera una definición formal. Lo que queremos sugerir aquí es que en la psicología de la creencia se puede hacer algo parecido: una teoría de la relación de confirmación entre, por ejemplo, perceptos visuales y perceptos lingüísticos podría postular *a)* un lenguaje neutral en que se puedan manifestar los dos; *b)* una forma canónica para estas representaciones, y *c)* principios computacionales que determinen el grado de confirmación de la oración en función, *inter alia,* de las relaciones formales entre su representación canónica y la representación canónica del input visual.

Existen, de hecho, pruebas empíricas de que al menos *alguna* información visual se «traduce» a un formato discursivo antes de ser utilizada para la confirmación o rechazo de las oraciones. (Véase, por ejemplo, Clark y Chase, 1972). Sin embargo, es importante insistir en que, aun cuando el caso de la traducción resulte tener carácter general, el problema de especificar los procedimientos para la comparación *directa* de las representaciones discursiva y no discursiva deberá ser abordado en algún punto dentro de una teoría psicológica de la fijación de la creencia. En términos aproximados podríamos decir que o bien habrá que determinar una relación de *confirmación* para los pares constituidos por preceptos visuales y representaciones discursivas, o bien, si es cierta la idea de la traducción, habrá que determinar también la relación de *traducción* que se da en esos pares. Anteriormente he dicho que los filósofos de la ciencia suelen limitarse a *presuponer* una representación canónica de los datos pertinentes para la confirmación de las teorías. Por lo general no se plantean la cuestión de cómo los datos entran en el lenguaje de datos. Pero los psicólogos tendrán que resolver la cuestión análoga si es que tratan de conseguir una explicación cabal de los procedimientos computacionales por los que las afirmaciones que oímos se comprueban con relación a lo que percibimos como hechos.

puesta el que las personas pierdan muchas veces información sobre el canal del input mientras que la conserven sobre lo que se comunicó por dicho canal. ¿*Leímos* que la suma de los ángulos de un triángulo equivale a 180 grados antes de *oírlo,* o viceversa? (Puede verse este tema desarrollado en Fodor, 1972. Rosenberg [1974] aporta datos experimentales y un modelo.)

Pero, independientemente de si las cosas son así o no, lo que interesa en este momento es que tienen que ser de *alguna* manera y que para ello hace falta que los mensajes caigan dentro del dominio de los principios, cualesquiera que sean, que determinen la relación de confirmación. Este requisito es conceptualmente independiente del requisito de que los mensajes constituyen inputs/outputs adecuados para los mecanismos que producen y analizan las instancias de oraciones. Podríamos imaginarnos una especie de organismo que fuera incapaz de utilizar lo que ve para comprobar lo que le dicen aunque, naturalmente, para este organismo el hecho de tener un lenguaje le sería mucho menos beneficioso que para nosotros. (Los pacientes con el «cerebro dividido» podrían aproximarse, en algunos aspectos, a estos organismos; cf. Sperry, 1956). Lo importante es que al incluir la teoría de la comunicación dentro de la teoría de la cognición, incrementamos las exigencias empíricas de cada una de ellas: por una parte, los mensajes deberán estar representados de tal manera que caigan dentro del dominio de la teoría de la fijación de la creencia, y, por la otra, los principios a los que apela dicha teoría deberán ser formulados de tal manera que se apliquen a objetos codificables lingüísticamente. Una psicología que cumpla esta doble exigencia está, ipso facto, mejor confirmada que la que sólo explique la codificación de los mensajes o sólo la confirmación de los mensajes que podamos codificar.

Pensemos en otro ejemplo. En el Capítulo 2 observamos que muchos filósofos actuales creen que para aprender un lenguaje natural hace falta (cuando menos) aprender una definición de verdad para el lenguaje. Se entiende que una definición de verdad es una teoría que empareja cada oración del lenguaje objeto, S_o con una oración metalingüística S_L, de tal manera que «$\ulcorner S_o \urcorner$ es verdad si y sólo si S_L» es en cuanto tal una consecuencia verdadera de la teoría semántica. Es de suponer que los filósofos que aceptan esta opinión lo hacen porque creen:

1. que comprender una emisión de una oración implica, en el menor de los casos, saber qué haría que la emisión fuera verdadera;
2. que una condición empíricamente necesaria para saber qué haría que fuera verdadera la emisión de una oración es computar una representación de la emisión que determine formalmente qué es lo que implica y qué es lo que la implica;
3. que una adecuada definición de verdad asociaría S_o con S_L únicamente si S_L determinara formalmente, en este sentido, lo que implica a S_o y lo que es implicado por S_o[11].

Dicho más económicamente, si alguien afirma, en cuanto semántico, que una teoría del significado empareja oraciones del lenguaje natural con fórmulas que repre-

[11] Para simplificar la exposición, me estoy comportando con gran tolerancia ante lo que debe figurar como vehículo de verdad e implicación y ante la relación tipo/caso en general. Lo que estoy diciendo podría decirse en forma mucho más rigurosa, pero para ello haría falta mucho más tiempo y espacio.

sentan sus condiciones de verdad, parece muy natural afirmar, en cuanto psicólogo, que la comprensión de una determinada emisión de una oración es cuestión de computar una fórmula que represente *sus* condiciones de verdad. El resultado es que la estructura cuya recuperación identificamos con la comprensión de la emisión de una oración debe ser un objeto adaptado formalmente por naturaleza a incluirse dentro de las reglas de inferencia que se aplican (informalmente) a la oración. Pero, por suposición, es la recuperación de los *mensajes* lo que constituye la comprensión de una oración pues, por suposición, lo que comunican las emisiones de oraciones son mensajes. Así, en la medida en que consideramos seriamente que las definiciones de verdad son teorías del significado, sabemos dos cosas sobre los mensajes: deben suministrar inputs/outputs adecuados para los modelos del hablante/oyente, y deben suministrar los dominios adecuados para las reglas de inferencia.

La idea de que una teoría del significado sirve, en efecto, para emparejar las oraciones del lenguaje natural con alguna forma de representación canónica de sus condiciones de verdad no es nueva, desde luego. Ha aparecido en las obras de filosofía desde el momento en que los filósofos comenzaron a distinguir entre la forma superficial de las oraciones y su forma «lógica». De hecho, la esencia de esta distinción ha sido señalar que las oraciones de un lenguaje natural no proporcionan dominios adecuados para la aplicación de las reglas lógicas, pero que sí lo harían ciertas traducciones especificables de estas oraciones. Representar la forma lógica de una oración es representar la condición de verdad de la oración de forma explícita, cosa que la propia oración no logra hacer.

Nuestra diferencia respecto a esta tradición es doble. En primer lugar, estamos considerando la noción de representación canónica en cuanto parte de una teoría *psicológica;* la representación canónica adecuada de una oración es la que el hablante tiene en mente cuando emite la oración y el oyente recupera cuando entiende lo que ha dicho el hablante; es decir, es aquella representación que hace explícito lo que intentan comunicar las emisiones de la oración. En segundo lugar, no hay ninguna razón concreta para que la representación se vea precisada únicamente a proporcionar un ámbito adecuado para las operaciones *lógicas.* Al fin y al cabo, existen procesos psicológicos distintos de la obtención de inferencias en los que interviene la información comunicada lingüísticamente, y, en la medida en que las representaciones canónicas contribuyen a la construcción teórica en psicología, sería conveniente que proporcionaran dominios adecuados también para esos procesos.

Para poner un último ejemplo, podemos decir que una de las cosas que podemos hacer con la información trasmitida lingüísticamente es olvidarla. Parece cierto que las diferentes partes de una oración no se olvidan al azar. Si yo os digo: «el chico y la chica fueron a la tienda» y más adelante os pido que me digáis lo que dije yo (es decir, emití), quizá os hayáis olvidado del chico o de la chica o del lugar al que fueron, pero no es posible que olvidéis el primer fono de «chico» y las palabras «a la». (Compárese con lo que ocurre cuando se intenta recordar cómo se llama alguien; aquí *es* probable que retengamos sólo la primera o dos primeras letras). Ahora bien, en el Capítulo 2 señalamos que el aspecto central del enfoque computacional en psicología es el intento de explicar las actitudes proposicionales del organismo por referencia a las relaciones que mantiene el organismo con las representaciones internas. Esta generalización vale, *inter alia,* para actitudes proposicionales como la de olvidar

que alguien ha dicho esto y aquello. En concreto, es posible construir la representación interna de lo que se ha dicho (o, para el caso, de cualquier otro percepto) mediante el requisito de que los elementos que se olviden juntos tengan una representación unitaria en el nivel de descripción para el que se definen los procesos de almacenamiento y recuperación. Lo que los psicolingüistas llaman «hipótesis de codificación» era, en realidad, un intento preliminar de determinar las representaciones de las oraciones que cumplen esta condición. (Véase una exposición más amplia en Fodor *et al.,* 1974). Evidentemente, la representación de una oración que proporcione un ámbito formal para los procesos de memoria, *y* exprese su significado, *y* suministre un input/output adecuado para un modelo del hablante/oyente, etc., habría dado razones suficientes para que se la reconociera como psicológicamente real.

Veamos el camino recorrido. Los objetivos teóricos de los lingüistas y psicolingüistas se pueden localizar convincentemente por medio de una teoría que considera la comunicación como la codificación y decodificación de mensajes. En la medida en que se pueda demostrar la realidad psicológica de las descripciones estructurales que proponen las gramáticas, podemos pensar que la lingüística describe el conjunto de representaciones computadas en el curso de este proceso de codificación/decodificación. De forma análoga, es plausible considerar que las teorías psicolingüísticas (supuestamente completas) especifican el orden en que se computan estas representaciones y los procesos de manipulación de la información que afectan a las computaciones. De esta manera, una teoría de la estructura de los mensajes se ve constreñida por una teoría de los lenguajes naturales, al menos en el sentido de que los mensajes deben proporcionar input/outputs adecuados para estos mecanismos computacionales.

Pero los mensajes deben especificar también la información que comunican las comunicaciones lingüísticas, y hemos visto cómo este requisito lleva consigo otros muchos. Podemos resumirlos diciendo que si un mensaje es aquella representación de una oración que es recuperada por alguien que entiende la oración que comunica el mensaje, en ese caso las operaciones cognitivas que están definidas en relación con la información que trasmiten las oraciones deben, ipso facto, estar definidas en relación con los mensajes. Si esto no es verdad, tendremos que abandonar el proyecto general de identificar los procesos cognitivos de los organismos con operaciones definidas en relación con las representaciones. Esta es, naturalmente, la consideración que relaciona lo que venimos diciendo de los lenguajes naturales con lo que dijimos antes sobre el lenguaje del pensamiento. Las fórmulas del código interno *son* precisamente aquellas representaciones sobre las que se definen las operaciones cognitivas; la única finalidad de suponer estas representaciones en los Capítulos 1 y 2 ha sido proporcionar ámbitos para las clases de procesos de manipulación de datos que proponen las teorías de la psicología cognitiva. Si los datos sobre el lenguaje y los procesos lingüísticos constriñen las teorías sobre los mensajes, también constreñirán las teorías sobre las fórmulas del lenguaje del pensamiento. Si el tipo de teoría de comunicación que he estado esbozando es correcto, los mensajes deben *estar* formulados en el lenguaje del pensamiento, es decir, deben ser fórmulas en cualquiera de los sistemas representacionales que proporcione los ámbitos para las operaciones cognitivas que se refieren (entre otras cosas) a la información transmitida lingüísticamente.

La mayor parte de lo que queda de capítulo va a ser un intento de hacer ver que

los hechos relacionados con los lenguajes condicionan realmente las teorías sobre los mensajes en la forma en que se indica en esta explicación. Antes de comenzar con ello, quiero hacer algunas puntualizaciones sobre la forma de entender la comunicación que he presentado. Creo que la importancia del tema justifica esta digresión. En primer lugar, en el Capítulo 2 indicamos que una característica de la organización de los ordenadores digitales de carácter general es que no se comunican en los lenguajes en que computan y que no computan en los lenguajes en que se comunican. La situación habitual es que la información entra y sale del código computacional mediante la intervención de sistemas de compilación que son, en realidad, algoritmos de traducción para los lenguajes de programación que «entiende» la máquina. Lo que nos interesa ahora es que, si es cierta la forma de entender la comunicación que estoy proponiendo, estas observaciones son válidas, con cierto detalle, para los mecanismos mediante los cuales los organismos intercambian información por medio de los lenguajes naturales. A todos los efectos, estos mecanismos constituyen «compiladores» que hacen posible que el hablante/oyente traduzca de las fórmulas del código computacional a las formas ondulatorias y viceversa[12]. Parafraseando una observación muy profunda que hizo en cierta ocasión el profesor Alvin Liberman (comunicación personal), parece claro, aunque no fuere más que por motivos biológicos, que los mecanismos de producción/percepción del lenguaje tienen un papel de mediación en la relación existente entre dos sistemas que les son muy anteriores en el tiempo: el aparato oído/boca que de hecho transduce señales verbales, y el sistema nervioso central que realiza las operaciones computacionales definidas sobre la información que comunican las verbalizaciones. El punto de vista a que nos referimos consistirá en afirmar que este proceso de *mediación* es fundamentalmente un proceso de *traducción;* es decir, traducción entre fórmulas de un lenguaje cuyos tipos describen formas ondulatorias y fórmulas de un lenguaje lo suficientemente rico como para representar los datos sobre los que actúan los procesos cognitivos. He mencionado antes

[12] La analogía entre los mecanismos psicológicos implicados en la comprensión de un lenguaje natural y los sistemas de complicación empleados para introducir y extraer la información de un computador digital se ha ido abriendo paso y son bastantes los teóricos que la aceptan. (Véase en especial Miller y Johnson-Laird, de pronta publicación). Sin embargo, no estoy suponiendo, como parecen hacer dichos autores, que la representación interna de una oración del lenguaje natural sea por lo general una «rutina» computacional (por ejemplo, una rutina para verificar la oración). Por el contrario, la representación interna de una oración es sencillamente su traducción al lenguaje del pensamiento; lo que viene a demostrar esto es que es perfectamente posible comprender lo que dice alguien sin tener la menor idea de cómo se puede verificar la afirmación que se ha hecho. Una afirmación no es normalmente una petición (o un mandato o ni siquiera una invitación) de averiguar si lo que afirma es cierto. Uno de los fallos que ha viciado gran parte del trabajo sobre simulación con máquinas de la comprensión de las oraciones ha sido no llegar a observar la distinción existente entre los procesos implicados en la comprensión de una elocución y los procesos implicados en su confirmación.

En pocas palabras, no vemos ninguna plausibilidad especial en la opinión, incorporada en lo que se denomina a veces «semántica procedimental», según la cual las oraciones del lenguaje natural suelen estar representadas por imperativos del código interno. Esta idea tiene su origen, en primer lugar, en el hecho de tomarse demasiado en serio el verificacionismo en cuanto doctrina sobre el significado, y, en segundo lugar, de interpretar demasiado literalmente la analogía hombre/ordenador como doctrina psicológica. En cierto sentido es verdad que los ordenadores reales se ocupan principalmente de los imperativos. Se podría decir que ello se debe a que su función característica es realizar las tareas que les preparamos. Pero las personas no tienen ninguna «función característica», y su interés por las oraciones consiste por lo general sencillamente en entenderlas.

que las teorías lingüísticas y psicolingüísticas, en la medida en que contribuyen a explicar la comunicación, deben especificar los procedimientos por los que se ve afectada esta traducción. Sin embargo, podríamos añadir, y con la misma justicia, que deben contribuir también a explicar hasta qué punto se *internalizan* los procedimientos en el curso de la adquisición del lenguaje. Imaginemos un mecanismo que se *aprenda* uno de sus compiladores y, si es correcta esta forma de entender la comunicación, estaremos imaginando un mecanismo que en algunos aspectos es como nosotros.

He dicho un mecanismo que se aprenda *uno* de sus compiladores, y esto nos hace pasar al segundo punto. Según la opinión que estamos exponiendo, existe una analogía bastante sorprendente entre los lenguajes naturales y las modalidades sensoriales. Evidentemente, existen procedimientos computacionales que reproducen una representación de las propiedades acústicas de un hecho de habla sobre una representación del mensaje que codifica. Pero es también evidente que éste no es el único sistema de que dispone el organismo para asociar las descripciones físicas de los inputs del entorno con las descripciones elaboradas en términos de variables cognitivamente pertinentes.

Supongamos que *F* es una fórmula del código interno que corresponde a la oración «En esta página hay un borrón» (de aquí en adelante, *S*). En ese caso es de suponer que la comprensión de las instancias de *S* implica asignar instancias de *F* como representaciones internas suyas, y que creer que cierta instancia de *S* es verdadera implica creer que también lo es la correspondiente instancia de *F*[13]. Una explicación natural de lo que supone creer que una instancia de *F* es verdadera es sencillamente que se *supone* que *F* es verdadera en aquellas computaciones en las que interviene; por ejemplo, que en dichas computaciones es tratada como un axioma no lógico.

Por eso, una forma de que *F* puede llegar a estar entre las fórmulas que se consideran verdaderas es que sea la fórmula que representa internamente una oración que se considera verdadera. Pero tiene que haber al menos otra forma; es decir, alguien ve esto: y cree lo que ve (es decir, considera que lo ve no es una alucinación, que es verídico, etc.).

Lo que yo afirmo es que hay algunas circunstancias en las que las consecuencias psicológicas de ver un borrón en la página son las mismas que las consecuencias psicológicas de leer que hay un borrón en la página; si creer lo que uno lee es condición suficiente para considerar que *F* es verdad, también lo es creer lo que ve. Esto tiene que ser así porque el borrón confirma lo que dice la frase, y parte de la comprensión de la frase consiste en comprender que esto es así. Todo esto es fácilmente inteligible bajo el punto de vista de que el estado computacional de un mecanismo que ve la mancha de tinta y entiende lo que ve es idéntico al estado computacional de un mecanismo que lee la frase y entiende lo que lee. Y además, no parece que sea inteligible como no sea desde ese punto de vista.

---

[13] Entre las sutilezas que estoy dispuesto a pasar por alto en este punto está el tratamiento de los elementos con subíndices. Si nos pusiéramos serios, habría que asegurarse de que *F* determine un referente concreto a «esta página». Una propuesta bastante común consiste en suponer que *F* contiene un esquema de una relación multi-local entre un hablante, una situación, un tiempo, etc.; los argumentos de esta relación serían distintos según las distintas instancias de *S*.

En pocas palabras, si el sistema de comprensión de oraciones funciona para reproducir los outputs del transductor en fórmulas del código interno, lo mismo ocurre con el sistema visual. Debemos partir de ello si queremos mantener, por una parte, que tener una creencia es cuestión de estar en una determinada relación computacional con una determinada fórmula interna y, por la otra, que las *mismas* creencias pueden depender de que se oigan oraciones y de que se vean borrones[14]. Por ello, si vamos a considerar que los mecanismos de percepción/producción de oraciones constituyen una especie de compilador, tenemos las mismas razones para considerar las modalidades sensoriales de la misma manera.

Volviendo a la analogía del ordenador: una de las razones por la que los ordenadores para usos diversos utilizan compiladores es precisamente que su utilización les permite realizar funciones diferenciadas. La información útil puede introducirse en la máquina en tantas formas diferentes como compiladores distintos tenga la misma pues, por el punto en que la información entra en los procesos computacionales, las diferencias en el código de input han quedado neutralizadas por las operaciones que realizan los compiladores. Tras la compilación, todos los inputs están representados por fórmulas del mismo lenguaje interno, por lo cual todas ellas están a disposición, al menos en principio, de todas las rutinas computacionales definidas en relación con las representaciones internas. Como señala Norman (1969, p. 164): «Una de las propiedades más importantes de los ordenadores es que no establecen ninguna distinción en su memoria entre instrucción, números y letras. Así, toda operación posible para el ordenador se puede realizar con cualquiera de los elementos almacenados».

Una vez más, según el planteamiento que estoy adoptando, la analogía entre personas y máquinas es bastante exacta. Las personas, como las máquinas, aceptan diferentes códigos de input, asegurando así una variedad de rutas a través de las cuales los procesos cognitivos pueden conseguir acceso a noticias relacionadas con el mundo exterior. Como en las máquinas, el truco consiste en tener compiladores para cada una de las modalidades de input. Los procedimientos de reconocimiento para los lenguajes naturales son uno de éstos.

A pesar de todo, existen muchas diferencias entre las personas y las máquinas (existentes). Una diferencia está en que las personas tienen más clases de sensores que cualquiera de las máquinas inventadas hasta la fecha; las personas pueden emparejar

---

[14] Convendría tener en cuenta que las cuestiones que estoy planteando aquí son diferentes, y en gran parte independientes, de la que hemos mencionado en la nota 10: si, en la confirmación de las oraciones mediante perceptos visuales, la relación de confirmación se define en relación con los datos visuales de forma directa o sólo en relación con sus traducciones a las fórmulas *discursivas* del código interno.

En efecto, podemos imaginarnos dos modelos de circulación de la información, cada uno de los cuales podría estar implicado en la confirmación visual de las oraciones. Si la «historia de la traducción» expuesta en la nota 10 es verdadera, la oración y el borrón visual se traducen en instancias del tipo *E,* y la confirmación de la oración por el percepto se consigue mediante el emparejamiento de identidad de estas instancias. Si la «historia de la traducción» es falsa, en ese caso la instancia de *F* que está directamente asociada con la oración se compara directamente con el input visual. En uno y otro caso, el organismo acaba por estar en la misma relación con las instancias de *F:* es decir, la relación de considerarlos como verdaderos. En resumen, las creencias que garantizan los inputs visuales deberán estar representadas por las mismas fórmulas que representan las creencias que garantizan los inputs lingüísticos, pues en muchas ocasiones se trata de las mismas creencias. Esto será cierto cualquiera que sea la opinión que se tenga sobre la forma en que las oraciones se ven confirmadas por los perceptos no lingüísticos.

las representaciones internas con más clases de representaciones físicas que las máquinas. Existen, de hecho, máquinas que pueden representar en su lenguaje de computación central la información transmitida visualmente. Dentro de ciertos límites estas máquinas pueden realizar sus computaciones sobre esta información y pueden integrarla con inputs que llegan a través de canales más convencionales (como fichas perforadas). Sin embargo, los inputs visuales que las máquinas actuales pueden compilar son *muy* rudimentarios, y la información que decodifican de las representaciones visuales es muy tosca en comparación con el sistema visual humano. Y no hay ninguna máquina que pueda, en este sentido, oler, gustar u oir.

La segunda diferencia es, naturalmente, que las personas pueden aprender nuevos procedimientos de compilación, es decir, pueden aprender nuevos lenguajes. Lo pueden hacer precisamente *porque* la relación entre instancias de oraciones y sus representaciones internas (a diferencia de la relación entre series visuales y *sus* representaciones internas) están mediadas por un sistema de convenciones. Pero si la capacidad de aprender estos sistemas de convenciones distingue al hombre de las *máquinas,* es justo añadir que le distingue también de todos los demás *organismos* (con el debido respeto a Sarah, Washoe, y el resto de sus colegas)[15].

Una reflexión final sobre el modelo de comunicación que venimos considerando. En el Capítulo 2 indiqué que era posible y provechoso pensar que un compilador que asocia cada una de las fórmulas del lenguaje de input *I* con alguna fórmula en el lenguaje de computación *C* es el metalenguaje en que se representan las propiedades semánticas de las oraciones de *I.* En efecto, la teoría del significado para las fórmulas de *I* es simplemente la función de traducción que las reproduce como fórmulas de *C.* En este contexto, sería plausible considerar que una teoría del significado para un lenguaje *natural* (como el inglés) es una función que conecta las oraciones de inglés con sus representaciones en el supuesto código interno.

Menciono este punto porque recientemente las «teorías de la traducción» del significado han sido objeto de bastantes críticas filosóficas, la mayoría de ellas, pienso yo, totalmente inmerecidas. Véanse, por ejemplo, las siguientes afirmaciones del profesor David Lewis.

Mis propuestas en relación con la naturaleza de los significados no obedecerán a las expectativas de aquellos lingüistas que consideran la interpretación semántica como la atribución a las oraciones y a sus componentes de «marcadores semánticos» y cosas semejantes (Katz y Postal [1964], por ejemplo). Los marcadores semánticos son *símbolos:* items del vocabulario de un lenguaje artificial que podríamos llamar *lenguaje de los marcadores semánticos.* La interpretación semántica hecha valiéndose de los mismos equivale simplemente a un algoritmo de traducción del lenguaje objeto al lenguaje auxiliar, el de los marcadores semánticos. Pero podemos saber la traducción al lenguaje de los marcadores semánticos de una oración inglesa sin saber lo más elemental sobre la significación de la oración inglesa: es decir, las condiciones en que sería

---

[15] Una de las ventajas de considerar las cosas de esta manera está en que aclara por qué *debe* haber universales lingüísticos. Para aprender un lenguaje natural, hay que aprender la correspondencia entre sus oraciones y sus representaciones internas. Pero es obvio que no podría haber una solución general para el problema de concebir un dispositivo que pueda aprender *cualquier* relación arbitraria entre los miembros de dos conjuntos infinitos. La posibilidad de construir un dispositivo capaz de aprender un lenguaje depende de que haya ciertas limitaciones en las clases de correspondencias que se le exigirá aprender.

verdadera. La semántica que no presente ningún tratamiento de las condiciones de verdad no es semántica. La traducción al lenguaje de los marcadores es, en el mejor de los casos, un sustituto de la semántica propiamente dicha, que se basa o en nuestra competencia tácita (en una fecha futura) como hablantes del lenguaje de los marcadores o en nuestra capacidad de elaborar una verdadera semántica al menos para el único lenguaje de los marcadores semánticos. También podrían servir las traducciones al latín, a no ser que los creadores del lenguaje de los marcadores decidieran incorporar en él caracteres prácticos —carencia de ambigüedad, gramática basada en la lógica simbólica— que permitieran conseguir una auténtica semántica más fácilmente que en el caso del latín... La semántica del lenguaje de los marcadores [no consiguió hacer frente a] las relaciones entre símbolos y el mundo de los no símbolos, es decir, a relaciones auténticamente semánticas (1972, pp. 169-170).

De momento, quiero mantener una apariencia de estricta imparcialidad en relación con los detalles de la teoría semántica propuesta en obras como las de Katz y Postal (1964). Volveremos a ocuparnos de estos temas en seguida y por extenso. Sin embargo, parece conveniente hacer algunas observaciones sobre la injusticia que se aprecia en las observaciones de Lewis si se consideran como una crítica general de los enfoques de la semántica basados en la idea de traducción.

Comenzando por lo que hay de acertado en las afirmaciones de Lewis, es cierto que la mera traducción de las fórmulas de *L* a las de un lenguaje canónico no constituye una explicación del modo en que las fórmulas de *L* se relacionan con el mundo. De ello se desprende que la semántica basada en la traducción, a diferencia de la semántica real o auténtica de que habla Lewis, no nos dice cómo se relacionan los símbolos con lo que simbolizan.

También es cierto que «podemos saber la traducción al lenguaje de los marcadores semánticos de una oración inglesa sin saber lo más esencial sobre la significación de la oración inglesa». Es cierto, pero está un poco fuera de lugar. Como la representación canónica de *S* es en cuanto tal una fórmula en algún lenguaje, es posible saber cuál *es* la representación canónica de *S* sin saber lo que *significa S:* por ejemplo, si no se entiende el lenguaje en que se formula la representación canónica. Pero, naturalmente, esto será válido para cualquier teoría semántica con tal que esté formulada en un sistema simbólico; y, por supuesto, no hay ninguna alternativa a formular así las propias teorías. Todos estamos en el mismo barco; todos tenemos que utilizar palabras para hablar. Como las palabras no tienen, por así decirlo, brillo propio como si fueran bolas de un árbol de Navidad, no hay forma posible de que una teoría semántica pueda garantizar que un determinado individuo va a encontrar sus fórmulas inteligibles.

Así, el sentido en que podemos «saber la traducción al lenguaje de los marcadores semánticos de una oración inglesa sin saber... las condiciones en que sería verdadera» no tiene mucho interés. Y lo que es sencillamente falso es que podamos *hacer* la traducción de una oración inglesa al lenguaje de los marcadores semánticos sin *representar* las condiciones en que es verdadera. Tenemos una garantía de que esto es falso en la misma definición de «traducción de *S* al lenguaje de los marcadores», pues ninguna fórmula satisface esa definición a no ser que sea verdad cuando, y sólo cuando, lo sea *S*[16].

[16] La verdadera diferencia entre la semántica real y la mera semántica de traducción no está en que solamente la primera proporcione una representación de las condiciones de verdad de las oraciones del len-

Finalmente, ¿hasta qué punto se puede admitir la crítica que dice: «Traducir el inglés a un lenguaje canónico no es mejor que traducir el inglés al latín, con la única diferencia de las posibles ventajas que el teórico pueda haber incorporado al primero y que Dios omitió cuando concibió el segundo»? Bien, como admite Lewis, las ventajas, para todos los objetivos prácticos, pueden ser esenciales incluso para hacer una semántica «real». Podría resultar perfectamente posible, por ejemplo, que no se pudiera describir la validez de los argumentos en inglés a no ser que antes se traduzcan a sus equivalentes canónicos. En realidad, parece seguro que es eso lo que ocurrirá de hecho, pues, en todas las versiones de que tengo noticia, lo menos que exigiría dicha descripción sería una notación libre de ambigüedades y, evidentemente, el inglés «superficial» no proporciona semejante notación.

Pero, en segundo lugar, la observación de que *T* es un «mero» esquema de traducción del inglés al latín no es probable que cause mucha impresión a una hablante que quiera saber lo que significa alguna que otra oración inglesa. Lo único que exige el caso es un mero esquema de traducción. Ahora bien, hemos supuesto que el sistema nervioso «habla» un lenguaje interno que no es ni inglés, ni latín, ni ninguna otra lengua humana. Las fórmulas de este código representan la información que transmiten las oraciones del lenguaje natural, por lo que una teoría que atribuya las fórmulas a las oraciones representa, ipso facto, los significados de las últimas. Y, aunque esta teoría no realice una semántica verdadera, en el sentido de Lewis, sin embargo debe ser interiorizada por cualquier organismo que pueda utilizar un lenguaje natural en cuanto vehículo para la comunicación. Pues sólo explotando las correspondencias que determina esta teoría pueden los organismos obtener la información que transmiten las verbalizaciones en una forma que puede ser utilizada por el sistema nervioso. En resumen, una no-realidad que goza de muy buena salud[17].

La primera mitad de este capítulo ha tratado de ofrecer una descripción general de la relación existente entre teorías lingüísticas y cognitivas; o, por decir lo mismo

---

guaje objeto; si *M* es la traducción de *S* al lenguaje de los marcadores, en este caso «*S* es verdad si y sólo si *M* es verdad» será una consecuencia lógica de la teoría semántica. La diferencia estriba, más bien, en la forma en que las dos clases de teoría caracterizan las propiedades semánticas de las expresiones del lenguaje objeto. En términos generales, las teorías de la traducción caracterizan tales propiedades haciendo referencia a las expresiones metalingüísticas que las comparten; las teorías semánticas «reales» no lo hacen.

Pensemos, por ejemplo, en la referencia misma. Las teorías de la traducción suelen especificar, para cada expresión referente del lenguaje objeto, una expresión correferente del metalenguaje. La referencia de las expresiones del lenguaje objeto está, por consiguiente, determinada, pero solamente en relación a una determinación de la referencia de las correspondientes expresiones del metalenguaje. Por otra parte, la semántica «real» *dice* de hecho a qué se refieren las expresiones del lenguaje objeto; es decir, nombra sus referentes. En consecuencia, la semántica «real» determina una relación de referencia, mientras que la «mera» semántica determina únicamente una relación de correferencia.

Lo que es cierto sin género de dudas es que una teoría de un lenguaje debe decir, de una u otra manera, a qué se refieren los términos del lenguaje. Por esta razón, una semántica «real» debería ser parte de una teoría del código interno. Naturalmente, esta consideración no hace que la especificación de un procedimiento de traducción de fórmulas del lenguaje natural a fórmulas del lenguaje interno sea una parte innecesaria de la teoría de aquél.

17 Además, si una teoría semántica «real» es aquella que indica cómo se relacionan con el mundo las fórmulas del código *interno,* los oyentes/hablantes *no* necesitan aprender semejante teoría; es de suponer que el código interno no se aprende sino que se da innatamente. (Véase el Capítulo 2).

en la modalidad material, de construir un modelo de la forma en que se interrelacionan los procesos lingüísticos y cognitivos. La intención de esta empresa era racionalizar el uso de los datos sobre el lenguaje para constreñir las teorías sobre la estructura del sistema representacional que media la cognición. Podemos resumir los resultados de la siguiente manera:

1. Las especificaciones de los mensajes representan la información que comunican las elocuciones de oraciones. Dicho de otra manera, representan la descripción según la cual el hablante intenta que se comprendan fundamentalmente sus verbalizaciones. O, dicho de una tercera forma, representan las intenciones comunicativas del hablante en la medida en que sus intenciones comunicativas se pueden interpretar (sólo) a partir de la forma de las palabras que emite.
2. La descripción de la correspondencia existente entre los mensajes y las formas lingüísticas que los expresan constituyen la misión propia de las teorías lingüísticas.
3. El hecho de que los hablantes oyentes pueden realizar estas correspondencias se debe explicar por la suposición de que han interiorizado los procedimintos computacionales que asocian instancias de mensajes con las instancias de oraciones y viceversa. La descripción del flujo de información que reproduce a través de tales procedimientos constituye la misión propia de las teorías psicolingüísticas.
4. Los mensajes deben estar representados de tal manera que proporcionen dominios adecuados para las computaciones implicadas en la codificación y decodificación del habla. La teoría de los mensajes se ve constreñida, por lo tanto, por la construcción de teorías psicolingüísticas.
5. Pero los mensajes deben estar representados de tal manera que proporcionen los dominios adecuados para procesos computacionales *no* lingüísticos en que incurre la información trasmitida verbalmente si, como afirma el punto 1, lo que transmiten las verbalizaciones son mensajes. La teoría de los mensajes se ve, pues, constreñida por la psicología cognitiva en general.

Algunas veces los filósofos dicen que las atribuciones de intenciones a las personas —especialmente las atribuciones de intenciones comunicativas a las personas— están tan fuertemente indeterminadas por los datos de la conducta que es engañoso describirlas como si fueran *empíricas*. Hay algo de verdad en esto, pero no demasiado. Lo que es probablemente cierto es que lo que *hace* cualquier organismo es compatible con una gran variedad de hipótesis sobre lo que pretende y, a fortiori, con un número indefinido de hipótesis sobre cómo se deben representar sus intenciones. Nuestra actitud, sin embargo, ha sido la de mantener que la principal constricción sobre las representaciones de las intenciones comunicativas no es la compatibilidad con la *conducta*, sino la compatibilidad con modelos razonables, e independientemente motivados, de los procesos psicológicos del hablante/oyente. Existe, naturalmente, un tipo de reduccionismo intransigente que supone que todas las constricciones de esta última clase deben resultar, a la larga, constricciones de la primera clase. Lo que no tenemos es ninguna razón plausible para pensar que lo que suponen estos reduccionistas acérrimos sea cierto.

Llegamos así a un punto decisivo de nuestra investigación. No basta con afirmar que la noción de lenguaje interno es conceptualmente coherente, que está impuesta por los modelos cognitivos adoptados en la actualidad por las personas sensatas, y que, en principio, las afirmaciones sobre la estructura de dicho lenguaje conectan con los temas empíricos de la psicología y lingüística. Lo que hay que demostrar ahora es que pueden hacerse algunos progresos en la valoración de estas afirmaciones. Este será el tema de las páginas que quedan. Voy a examinar algunos tipos de argumentos que son muy conocidos en lingüística (este capítulo) y psicología (Capítulo 4), pero interpretaré estos argumentos de forma un tanto excéntrica, es decir, en cuanto que repercuten sobre los temas relacionados con el carácter del código interno.

Un tipo de pregunta que suele ser conveniente hacer al hablar de un sistema representacional es la siguiente: ¿cuáles son los items de su vocabulario? No hay, naturalmente, ninguna *garantía* de que este tipo de pregunta vaya a tener sentido desde el momento en que ciertos sistemas representacionales no tienen vocabularios (suponiendo que tener un vocabulario es tener un inventario finito de items elementales, discretos y significativos); piénsese en los sistemas representacionales «analógicos» como el lenguaje de las abejas o el de las imágenes. Sin embargo, a primera vista, es razonable suponer que lo tenga un sistema lo suficientemente rico como para expresar los mensajes que pueden transmitir las oraciones del lenguaje natural. En cualquier caso partiremos de esa suposición a efectos heurísticos y consideraremos algunas de las pruebas lingüísticas sobre el vocabulario del lenguaje del pensamiento.

EL VOCABULARIO DE LAS REPRESENTACIONES INTERNAS

Tradicionalmente se ha afirmado que las oraciones de los lenguajes naturales podrían comunicar lo que comunican aun en el caso de que sus vocabularios fueran más reducidos de lo que son en realidad. La idea de fondo es que es posible «eliminar» algunos items del vocabulario del lenguaje natural definiéndolos por medio de otros, manteniendo al menos el conjunto de inferencias que se pueden deducir válidamente de las oraciones del lenguaje. Supongamos, por ejemplo, que «soltero» significa lo mismo que «hombre no-casado». En ese caso, lo que se pueda decir en un lenguaje que contenga a ambos se puede decir en un lenguaje que contenga solamente a uno de los dos. Además, si vamos a prescindir de uno de ellos, es claro cuál tendrá que ser: los componentes «hombre» y «no-casado» aparecen en frases distintas de la mencionada, y no tendría mucho sentido eliminar la expresión si no podemos eliminar sus componentes. En resumen, si eliminamos «hombre no-casado» en favor de «soltero», no hemos reducido el número de items del vocabulario del lenguaje, aunque hemos reducido el número de locuciones. Pero si procedemos a la inversa, considerando «soltero» como la expresión definida, el lenguaje puede arreglárselas con sólo dos elementos primitivos mientras que en el caso anterior había tres. Si aplicamos este tipo de argumentación en todos los casos donde sea posible, llegamos a la idea de la *base primitiva* del vocabulario del lenguaje, es decir, el conjunto más pe-

queño de items del vocabulario en función de los cuales se puede determinar todo el vocabulario[18].

El interés que representa para nosotros la idea de una base primitiva es el siguiente: hemos visto que el sistema de representaciones internas para las oraciones de un lenguaje natural debe captar al menos la capacidad expresiva de dichas oraciones. Ahora bien, parece que la base primitiva de un lenguaje determina su capacidad expresiva en la medida en que esta última está en función del vocabulario. Por eso, es una posibilidad abierta el que el vocabulario del sistema utilizado para representar los mensajes transmitidos por las oraciones de un lenguaje natural corresponda precisamente a la base primitiva de ese lenguaje. Si fuera este el caso, se deduciría, por ejemplo, que «Es soltero» y «Es un hombre no casado» reciben representaciones idénticas en cuanto al mensaje, suponiendo que «soltero» y «hombre no casado» sean sinónimos.

Por supuesto, del hecho de que el vocabulario primitivo del sistema representacional interno *pueda ser* más pequeño que el vocabulario «superficial» de un lenguaje natural, no se deduce que el vocabulario primitivo de ese sistema *sea* más reducido que el vocabulario superficial de un lenguaje natural. Siempre cabe la posibilidad de que el vocabulario del lenguaje interno sea más rico de lo que necesita: es decir, más rico de lo que hace falta a efectos de expresar el contenido de las oraciones del lenguaje natural. Eso es, supongo yo, una cuestión *rigurosamente* empírica, y es la cuestión de la que nos ocuparemos fundamentalmente a continuación.

En las publicaciones lingüísticas aparecidas recientemente parece que hay un consenso considerable en el sentido de que existe un nivel «semántico» de representación gramatical —un nivel en el que se especifica formalmente el significado de las oraciones— y que, cualesquiera que sean las otras propiedades que pueda tener este nivel, es al menos claro que proporciona representaciones idénticas para las oraciones sinónimas. Como normalmente se suele defender la realidad psicológica de las estructuras que enumeran las gramáticas, y como lo que representan las representaciones semánticas son mensajes, esto equivale, en nuestra perspectiva, a la afirmación de que los mensajes *están* formulados en un vocabulario menos rico que el vocabulario superficial de los lenguajes naturales. Hay muchos lingüistas que afirman que en el curso de las derivaciones gramaticales se produce un proceso análogo a la sustitución del *definiendum* por el *definiens,* y que hay varios hechos semánticos y/o sintácticos relacionados con el lenguaje natural que se explican por la existencia de este proceso. Personalmente estoy convencido de que este consenso no está justificado, como trataré de explicar poco más adelante. Antes de nada, quiero explorar con cierto detalle las clases de mecanismos que han propuesto los lingüistas para conseguir el resultado de la definición eliminatoria, y los argumentos en favor y en contra de la suposición de estos mecanismos.

Quizá la primera obra donde se trató este conjunto de temas dentro del contexto de la gramátiea generativa fue la de Katz y Fodor (1963). La proposición básica propuesta en este trabajo no ha sido casi objeto de ataques por parte de los gramáticos

[18] En beneficio de la sencillez, voy a suponer que existe exactamente ese conjunto. Evidentemente, no debe haber más de un conjunto de items de vocabulario *psicológicamente* primitivos para que se pueda hacer psicolingüística.

generativos que aceptan una visión «interpretativa» de la semántica: es decir, que uno de los mecanismos computacionales que median la relación entre representaciones semánticas y oraciones superficiales es un *diccionario* y que una de las cosas que el diccionario dice sobre el inglés es que «bachelor» (=soltero) corresponde a la fórmula metalingüística *unmarried man* (= hombre no casado)[19]. Por supuesto, esto es considerar muy literalmente la noción de *definición* eliminatoria. El nivel semántico proporciona la misma representación para las oraciones superficiales «Es soltero» y «Es un hombre no casado». Además, considera que la segunda es más explícita, pues la representación que suministra en ambos casos será un compuesto de las representaciones que suministra para «No está casado» y «Es un hombre». En efecto, el nivel semántico prescinde de la diferencia entre «soltero» y «hombre no casado» pero es sensible al hecho de que la última expresión tiene como componentes a «hombre» y «no casado». Por lo tanto, si un hablante quiere introducir «soltero» en una oración superficial, o si un oyente quiere entenderlo, deben hacerlo a través de su conocimiento del diccionario. Según esta explicación, no hay ningún item del vocabulario de la representación semántica que corresponde a «soltero» a no ser los items que corresponden directamente a «hombre no casado».

A pesar de las desventajas que pudiera presentar una teoría de este tipo, al menos tiene en consideración los siguientes tipos de datos:

1. «He is a bachelor» (= es soltero) y «He is an unmarried man» (= es un hombre no casado) son sinónimos, es decir, son expresiones alternativas del mismo mensaje.
2. «Bachelor» (= soltero) debe definirse en función de «unmarried man» (= hombre no casado), y no viceversa.
3. Lo que se deduce de «He is a bachelor» se deduce también de «He is an unamarried man» y viceversa. (Esto se verá confirmado por la suposición de que las operaciones inferenciales son sensibles sólo a la representación del *mensaje* de una oración; es decir, los dominios en que se aplican las reglas de inferencia son representaciones semánticas más que formas superficiales. Como «He is a bachelor» y «He is an unmarried man» tienen, por suposición, la *misma* representación semántica, se deduce que una inferencia será representada como válida en un caso si es representada como válida en el otro.)

La objeción más seria que las obras publicadas han presentado a este tipo de teoría es, pienso yo, simplemente que es demasiado liberal. Decir que la gramática contiene un diccionario es, al fin y al cabo, decir únicamente que contiene un conjunto finito de parejas cada una de las cuales estaría compuesta por un término definido del lenguaje natural junto con su fórmula definidora en el sistema representacional. La dificultad está en que, a no ser que tengamos una información anterior sobre cuáles son las fórmulas bien elaboradas del sistema representacional, la propuesta interpretativista haría posible que prácticamente *todo* pudiera figurar como posible defi-

---

[19] Cuando la distinción sea importante, utilizaré comillas cuando se trate de expresiones del lenguaje natural y letra cursiva para las expresiones del vocabulario de las representaciones semánticas. Por eso, estrictamente hablando, la afirmación a que nos estamos refiriendo no es que «bachelor» se define en función de «unmarried man», sino que «bachelor» y «unmarried man» se definen ambos en función de *unmarried man.*

nición, y esto no puede ser cierto. Tiene que haber ciertas limitaciones sobre lo que puede ser una expresión definidora, pues tiene que haber ciertos límites en lo que puede significar una palabra. Por ejemplo, *of and but* (= de y pero) no es una definición posible (no es el significado de una posible palabra) porque, dicho sin rodeos, *of and but* no significa nada. Pero ¿qué hay en la teoría semántica descrita hasta ahora que descarte a *of and but* como *definiens*? En efecto, la indicación de que la interpretación semántica implica la aplicación de un diccionario resulta prácticamente vacía de contenido a no ser que se pueda señalar algo que constriña lo que puede aparecer *en* el diccionario; es decir, algo que especifique la forma y contenido de las posibles definiciones. (Cfr. Katz, 1972, especialmente el Capítulo 3, donde se presenta un intento sistemático de formular estas constricciones dentro de los supuestos de la semántica «interpretativa».)

Veamos el camino recorrido: podríamos calcular un límite inferior en el tamaño del vocabulario del nivel del masaje ni supiéramos cuáles son las expresiones del lenguaje natural que están reemplazadas por definiciones en el curso de la computación de representaciones realizadas en ese nivel. Podríamos decir algo sobre cuáles son las expresiones que se ven reemplazadas por definiciones si supiéramos cuáles son las definiciones y, en concreto, cuáles son las constricciones sobre las expresiones definidoras. Ahora bien, lo que sabemos es esto: cualesquiera que sean las fórmulas del lenguaje del mensaje que expresan las definiciones deben estar al menos bien construidas en ese lenguaje. La estrategia de búsqueda recomendada consiste, por tanto, en utilizar lo que se pueda averiguar sobre las condiciones de corrección en el lenguaje del mensaje para constreñir la clase de posibles definiciones y utilizar lo que se puede averiguar sobre las constricciones de las posibles definiciones para iluminar las condiciones de corrección en el lenguaje del mensaje. Esto es lo que viene ocurriendo (más o menos explícitamente) en la semántica lingüística desde hace varios años. (Janet Dean Fodor está a punto de publicar un amplio estudio de este tipo de trabajos.)

Lo primero que hay que tener en cuenta es que las sugerencias más obvias no funcionan. Consideremos, por ejemplo, la posibilidad de que la corrección del lenguaje del mensaje cumpla con las condiciones de corrección superficial en cualquiera de los lenguajes naturales que se utilicen para expresar los mensajes. Esto significaría, en concreto, que las definiciones correspondientes a las palabras de *L* tienen la sintaxis de las fórmulas correctas de *L,* de manera que, al menos en lo que se refiere a los imperativos sintácticos, toda palabra de *L* que se pueda definir de alguna manera puede ser definida por alguna expresión de *L*.

Como es lógico, no hay una razón a priori por la que las condiciones de corrección de las fórmulas del lenguaje interno deban reflejar las condiciones de corrección de las oraciones superficiales. Por el contrario, si, como hemos supuesto, el lenguaje del pensamiento es un sistema distinto de los lenguajes naturales, las correspondencias entre sus estructuras deberían considerarse como hechos *sorprendentes,* hechos que necesitarían una explicación. En concreto, no hay ninguna razón a priori por la que las definiciones de los términos de *L* deban ser expresables en *L*. Insisto en ello porque se puede discutir que sea *posible* expresar de esta manera algunas clases de definiciones comunes. Por ejemplo, si es cierto que «dog» (= perro) significa *domestic canine* (= canino doméstico), esa virtud se puede expresar mediante una fórmula

gramatical del inglés (es decir, «"dog" means "domestic canine"») [«means» = significa]. «Domestic canine» es un nombre inglés bien construido, por lo que el inglés superficial puede servir como metalenguaje propio en orden a expresar esta parte de la semántica inglesa.

En realidad, podríamos generalizar esta observación. Las definiciones de los nombres léxicos (si tienen realmente definiciones) se pueden expresar generalmente mediante frases bien construidas con la estructura (adjetivo + nombre). Y haciendo una extrapolación, podríamos sentirnos tentados de decir: las definiciones de los términos de *L* deben admitir la posibilidad de ser formuladas como componentes superficiales de *L*. Sin embargo, es probable que esto sea un error.

Los problemas serios se presentan en el caso de lo que podemos denominar ampliamente como expresiones «relacionales». Pensemos, por ejemplo, en «or» (= o). Si alguien tratara de mantener que es posible definir «or» cuando está solo, la fórmula definidora tendría que ser probablemente algo parecido a «not both ((not...) and (not...))» [= no ambos ((no...) y (no...))]; y, sea lo que sea esa fórmula, no es una secuencia bien construida en el inglés superficial. Son parecidas las dificultades que se presentan con los verbos relacionales como «kill» [= matar], del que algunos piensan que significa *cause to die* [= hacer morir]. Téngase en cuenta que «cause to die» no puede darse como componente de una oración inglesa: «John caused to die Bill» no está bien construida[20].

Los filósofos interesados en la definición eliminatoria (pero no, en general, en la realidad psicológica o en la plausibilidad lingüística) han tratado generalmente estos casos, al menos desde Rusell (1905), recurriendo a las «definiciones por el uso». Pero de esa manera no se define de hecho «or» o «kill». Más bien lo que se hace es introducir reglas para eliminar «*P* or *Q*» en favor de «not both ((not *P*) and (not *Q*)) [= no ambos ((no *P*) y (no *Q*))], y «*x* killed *y*» [= *x* mató a *y*] en favor de «*x* caused *y* to die» [= *x* hizo morir a *y*]. Así, donde las definiciones convencionales relacionan *palabras* definidas y *frases* definidoras, la definición «por el uso» relaciona *frases* con frases. Como las frases así definidas pueden contener variables que pueden, por así decirlo, ser traspasadas a las expresiones definidoras, es muy probable que un empleo sistemático de las definiciones por el uso permita cumplir la condición de que todas las definiciones deben ser constituyentes superficiales bien construidos. Lo que es seguro es que esto nos permitirá estar más cerca de cumplir esa condición que la utilización de definiciones convencionales.

Todo esto va muy bien para los objetivos que han tenido presentes los filósofos —es decir, simplificar la base primitiva del lenguaje—, con tal que sea posible agotar de modo finito los contextos sintácticos en que ocurre el «eliminandum» («or»,

---

[20] «Cause to die» se reconoce inmediatamente como «lenguaje de diccionario», lo que equivale a decir que, por lo general, los diccionarios no cumplen la condición de que las definiciones deben estar formuladas en la síntaxis del inglés superficial. Naturalmente, muchas entradas que aparecen en los diccionarios no son definiciones en absoluto, suponiendo que las definiciones son frases sinónimas a los términos que definen. Fund y Wagnalls (1966) tratan la palabra «or» [= o] enumerando sus usos, no diciendo lo que significa: «1. Presentando una alternativa: pararse *o* irse... 2. Ofreciendo una selección dentro de una serie: ¿Vas a tomar leche *o* café *o* chocolate?...». De hecho, una de las conclusiones de este capítulo será que la importancia de la relación de definición se ha supervalorado enormemente en las obras de semántica lingüística.

«kill», o lo que sea). Sin embargo, si se considera como una explicación psicológica, o de lingüística descriptiva, resulta poco convincente. En efecto, las definiciones por el uso pueden *reforzar* las constricciones sobre las expresiones *definidoras* (insistir en que deben ser constituyentes) precisamente porque *relajan* las constricciones sobre las expresiones *definidas* (admiten la posibilidad de que las exigencias definidas sean frases y no palabras). O, dicho de otra manera, el recurso a las definiciones por el uso va contra la intuición fuertemente arraigada de que, si prescindimos de los modismos, las expresiones definibles de un lenguaje proceden todas del mismo nivel lingüístico (es decir, palabras o morfemas). En realidad, el problema de la definición por el uso es que trataría «*P* or *Q*» [*P* o *Q*] de forma parecida a como trata «kick the bucket» [= estirar la pata]; es decir, como una cadena que tiene una estructura sintáctica interna pero que no se puede descomponer en elementos semánticos con significados especificables por separado. Hasta hace poco los filósofos han sido bastante generosos con la pérdida de estructura lingüísticamente significativa con tal de que ello estuviera en conformidad con el carácter finito de las teorías semánticas que patrocinaban. Pero los lingüistas y psicolingüistas no pueden ser tan tolerantes; no se preocupan únicamente del aspecto formal sino también de la verdad empírica, y la verdad empírica parece ser que «*P* or *Q*» no es una forma de frase hecha.

Entre las aportaciones más interesantes a esta maraña de problemas relacionados con las definiciones y las constricciones sobre las definiciones, se puede mencionar un grupo de propuestas asociadas al epíteto «semántica generativa». La idea básica es que sería posible considerar las definiciones como especies de relaciones *sintácticas;* en concreto, que los términos definidos podrían derivarse de sus expresiones definidoras mediante reglas formalmente indistinguibles de las transformaciones sintácticas[21]. Si esto es cierto, sería posible utilizar la sintaxis del lenguaje objeto para constreñir las posibles definiciones de sus términos. Pues aunque no sea posible admitir que toda definición deba ser un constituyente superficial bien formado del lenguaje objeto, sería necesario que toda definición fuera el output de un proceso sintáctico del lenguaje objeto. En concreto, sería necesario que toda definición fuera una fórmula bien formada en *algún* punto de una derivación sintáctica. Dicho en términos más generales, para que esta propuesta sea válida, algunas de las constricciones sobre las definiciones deberían ser «heredadas» de las constricciones sobre las transformaciones sintácticas.

Por ejemplo, suele aceptarse generalmente como una constricción de las transformaciones el que sólo los constituyentes puedan moverse, suprimirse o sustituirse; es decir, los objetos a los que se aplica una transformación deben ser constituyentes en el punto de una derivación en que se les aplica la transformación. Este requisito (al que de aquí en adelante denominaremos «single node constraint» [= constricción del nodo único] o SNC)[22] está profundamente arraigado en la teoría generativa, pues una forma clásica de demostrar que algo *es* un constituyente en un punto determinado de

[21] Estrictamente hablando, lo que se quiere decir no es que los términos procedan de su definición (del lenguaje objeto), sino que ambos derivan de una fuente (metalingüística) común; es decir, «kill» y «cause to die» derivan de *cause to die.* (Véase nota 19, más arriba).

[22] Estrictamente hablando, el principio en cuestión es que las transformaciones elementales sólo se pueden aplicar a subárboles de un árbol de una estructura constituyente. Una colección de nodos constituye un subárbol si hay un nodo del superárbol que los domina a ellos y sólo a ellos.

una derivación es demostrar que cierta transformación lo mueve, suprime o sustituye en ese punto. Pensemos por ejemplo en el par de oraciones (1) y (2).

(1) Bill climbed over the fence.
(Bill saltó por encima de la cerca)

(2) Bill phoned up the man.
(Bill telefoneó al hombre)

Generalmente se acepta que en la primera frase las palabras deberían ir agrupadas así: (Bill) (climbed) (over the fence), mientras que en la segunda los grupos serían: (Bill) (phoned up) (the man). Es decir, «phone up» es un constituyente de (2), pero «climb over» *no* lo es de (1). El argumento en que se basa esta afirmación hace relación directa al SNC; es decir, que hay una oración (3) que corresponde a (1), pero no hay una oración (4) que corresponda a (2).

(3) Over the fence climbed Bill.

(4) *Up the man phoned Bill.

Como hay una transformación que se aplica a «over the fence», el SNC exige que tal secuencia sea considerada como un constituyente.

Hasta ahora hemos visto lo siguiente: si los términos definidos son sintácticamente derivables de sus expresiones definidoras (cfr. nota 19), las definiciones deberán cumplir todas las constricciones que se aplican a los objetos que caen en el dominio de las transformaciones. En concreto, deberán acomodarse a SNC. Pero SNC exige que los objetos a que se aplican las transformaciones sean constituyentes en el punto en que se aplican las transformaciones. Por eso, si la explicación de la semántica generativa es cierta, sabemos sobre las constricciones de las definiciones al menos lo siguiente: las expresiones definidoras deben ser constituyentes correctamente construidos en uno u otro punto en el curso de las derivaciones sintácticas del lenguaje objeto. Esta constricción, lógicamente, es más débil que la exigencia de que las definiciones sean constituyentes *superficiales* bien construidos del lenguaje objeto; pero es, sin embargo, lo suficientemente fuerte como para resultar interesante.

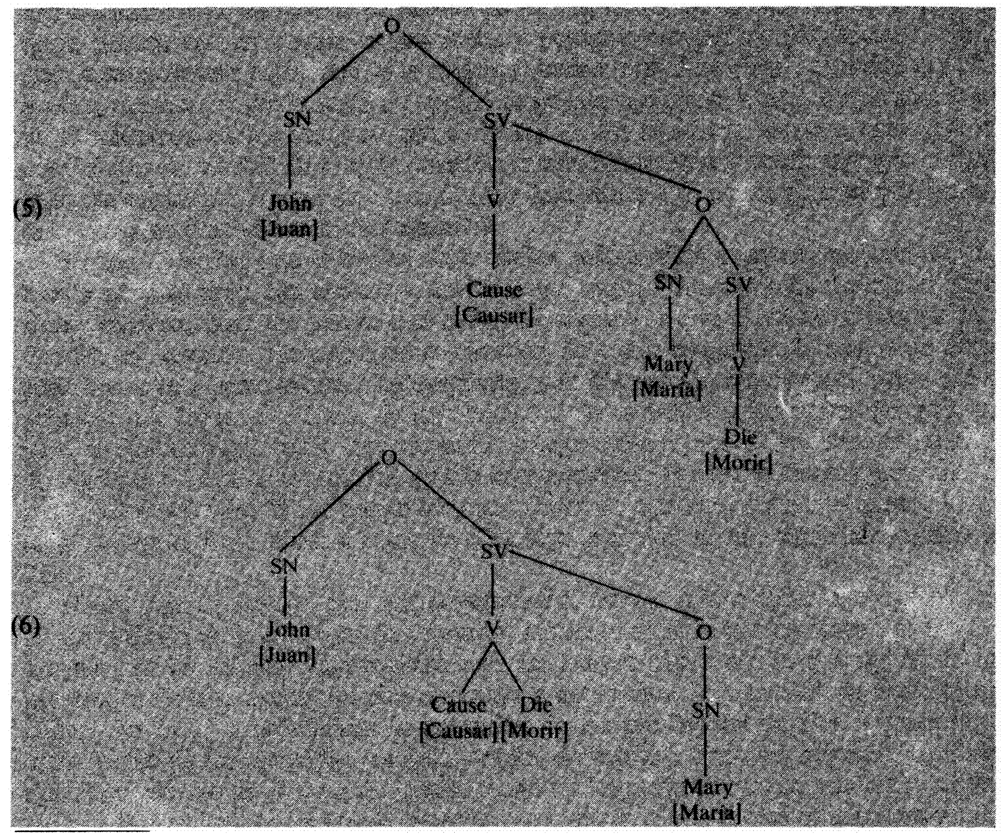
Pensemos ahora cómo podría funcionar la presente propuesta con relaciones definitorias como las que puede haber entre «kill» [= matar] y «cause to die» [= hacer morir][23]. Empezaríamos suponiendo que (5) está entre las estructuras profundas sin-

---

[23] Lo que sigue es una versión simplificada del tratamiento de las causativas propuesto por Lakoff (1970a) y McCawley (1971) entre otros. Puede verse un desarrollo detallado de algunas de las dificultades de este tratamiento en Fodor (1970), pero vale la pena insistir desde el primer momento en que la definición propuesta es sin duda deficiente, pues «*x* caused y to die» no implica «*x* killed y». Pensemos el caso en que *x* hace morir a *y* consiguiendo que lo mate ***algún otro***.

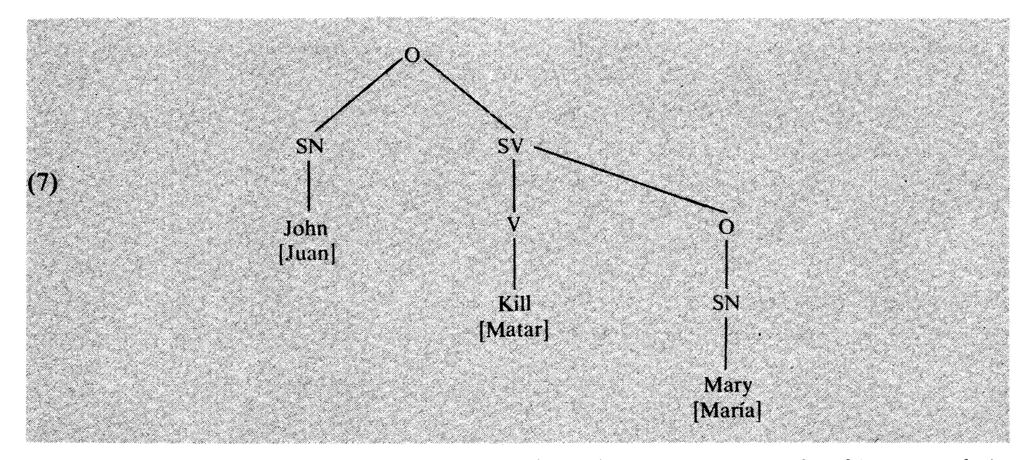
Lo habitual es responder a este tipo de objeción invocando una relación especial de «causalidad inmediata» de tal manera que, por decreto, «*x* inmediately caused *y* to die» implica «*x* killed *y*». Es esta relación de causalidad inmediata la que se dice que figura en definición de verbos como «kill». Es un misterio (en apariencia un misterio que va a permanecer sin explicación de forma permanente) en qué consiste,

tácticas correctas del inglés. (La estructura (5) es la que, en el sentido más obvio, está directamente en la base de oraciones como «John caused Mary to die», «John caused Mary's death» [= John hizo morir a Mary, John, causó la muerte de Mary], etcétera). Tendremos que suponer dos transformaciones. La transformación denominada *predicate raising* [= subida del predicado] se aplica al verbo de la oración incrustada, con el resultado de unirlo al nodo verbal de la oración principal. La aplicación de la transformación *predicate raising* a (5) da lugar a una estructura derivada, como (6).



exactamente, esta relación. (En el sentido más obvio de «inmediately cause» lo que causa inmediatamente la muerte no es, generalmente, lo que mata. Si lo fuera, todos moriríamos de un paro cardiaco). Pero cualquiera que sea la noción de causalidad inmediata a que se llegue, la respuesta está al margen del tema. Lo que importa es que, de todas las especies de *x* que hacen que *y* muera, hay una y sólo una que es necesaria y suficiente para conseguir que sea verdad «*x* killed *y*»: es decir, que *x* causara la muerte de *y* matando a *y*. De la misma manera, de todas las especies de *x* que hacen que el vaso se rompa, existe una y sólo una que es necesaria y suficiente para conseguir que sea verdad «*x* rompió el vaso»: es decir, que *x* hiciera que se rompiera el vaso rompiendo el vaso. Y así sucesivamente, *mutatis mutandis,* para el resto de los verbos causativos. Supongo que esto es una indicación poderosa de que tanto «kill» como «cause to die» (tanto «$\text{romper}_{\text{transitivo}}$» como «hacer que se $\text{rompa}_{\text{intransitivo}}$») deben ser items del vocabulario de un metalenguaje lo suficientemente rico como para representar las condiciones de verdad de las oraciones inglesas. Más en concreto, es una indicación clara de que SN_1 debe estar representado en cuanto agente de $V_{\text{causativo}}$ (y no en cuanto agente de $causar_v$) en el análisis semántico de oraciones de la forma superficial SN_1 $V_{\text{causativo}}$ SN_2.

Convendría tener en cuenta que en (6) «cause die» se analiza como un verbo compuesto; por eso especialmente «cause die» cumple con SNC y es, en esa medida, un posible ámbito de nuevas transformaciones. Y, de hecho, podrá aplicarse ahora una nueva transformación. La *lexicalización* es una transformación por sustitución que convierte estructuras como (6) en estructuras como (7); es decir, en árboles superficiales que contienen términos definidos.



Lo importante a tener en cuenta sobre la *lexicalización* es que su función es en algún sentido análoga a una definición de «kill» [=matar] y no a un definición por el uso de «*x* kill *y*». Lo que pone a «kill» en relación correcta con su sujeto y objeto en oraciones superficiales como «John killed Mary» [= John mató a Mary] no es que esté definido en el contexto de variables (como ocurriría en el caso de la definición por el uso); más bien, la derivación está dipuesta de tal manera que, después de la *lexicalización,* el sujeto profundo de «cause» [= causar, hacer] se ha convertido en el sujeto derivado de «kill», y el sujeto profundo de «die» [= morir] se ha convertido en el objeto derivado de «kill».

Una propuesta como ésta nos invita a plantear tres tipos de pregunta: ¿es deseable, es técnicamente posible y hay prueba en su favor? Creo que no puede caber ninguna duda de que la respuesta a la primera pregunta debe ser afirmativa. Lo que hemos echado en falta todo el tiempo ha sido una forma de constreñir las posibles definiciones para poder llegar a calcular la riqueza de la base primitiva del sistema de representaciones semánticas. Si la propuesta que nos ocupa fuera correcta, nos proporcionaría esta fuente de constricciones: sabríamos al menos tanto sobre las condiciones de las definiciones como sobre las condiciones de las transformaciones. Además, si hay analogía entre las constricciones sobre las condiciones de buena formación en el sistema representacional y las constricciones sobre las condiciones de buena formación en el lenguaje objeto, se pueden justificar con esta explicación: las reglas que relacionan las palabras con sus definiciones son un caso especial de las reglas que relacionan las estructuras superficiales con las estructuras profundas.

A continuación voy a ocuparme brevemente de la cuestión de la viabilidad técnica. Me parece falso que los buenos candidatos para las definiciones cumplan invariablemente con las constricciones sobre transformaciones como la SNC, lo mismo que

parece falso que los buenos candidatos para definiciones constituyan invariablemente frases superficiales bien formadas. Sin embargo, en este momento, quiero echar una ojeada a la tercera pregunta. Una de las afirmaciones fundamentales de los semánticos generativos ha sido que el suponer que existen relaciones sintácticas entre *definiendum* y *definiens* nos permite explicar una gran variedad de hechos gramaticales que no se pueden explicar de otra manera. Si esto es cierto, su importancia es evidente, pues constituye una base empírica para el tratamiento sintáctico de las definiciones. Constituiría un ejemplo muy adecuado del uso de los datos lingüísticos distribucionales para elegir entre las teorías relacionadas con las representaciones internas, y la idea central de este capítulo es que los datos lingüísticos *puedan* elegir entre estas teorías. Por eso, vista desde esta perspectiva, la propuesta de la semántica generativa es importante aunque *no* sea cierta, con tal de que haya datos que demuestren que no lo es. Nuestra afirmación principal es que las teorías sobre las representaciones internas son teorías empíricas legítimas. Una forma de demostrar que lo son es encontrar datos que las confirmen. Pero también sería positivo demostrar que hay datos con las que son incompatibles.

No entra dentro de los objetivos de este libro intentar ofrecer un examen detallado de las pruebas a favor y en contra de un tratamiento sintáctico de las definiciones. Lo que voy a hacer es estudiar un ejemplo. En concreto, quiero hacer ver cómo algunos datos sobre las oraciones inglesas se podrían resolver mediante una cualquiera de entre tres suposiciones diferentes sobre el carácter y contenido del vocabulario primitivo del sistema de representaciones subyacentes. Mi conclusión será que, al menos en relación con estos datos, la mejor de las soluciones es la que supone no sólo que no hay ningún proceso *sintáctico* de definición, sino que no hay ninguna clase de proceso para la definición; es decir, que tanto la expresión definida como su definición aparecen como items en el vocabulario primitivo del sistema representacional.

Ni que decir tiene que esta forma de argumentación no puede refutar la opción de la semántica generativa. Si este ejemplo no confirma la existencia de un proceso transformacional de *lexicalización,* quizá lo haga otro ejemplo. Sin embargo, mi objetivo es sencillamente presentar algunos ejemplos de las clases de consideraciones que son pertinentes, y no aprobar o condenar un tratamiento determinado de la definición. Por una parte, veremos que la explicación sintáctica de las definiciones parece no funcionar al menos en un caso en el que, a primera vista, sería de esperar que lo hiciera; y, por la otra, descubriremos algunos hechos que parecen ser informativos en relación con el carácter de las representaciones semánticas independientes de la forma de entender la definición que se acepte. Ya al final, examinaré ciertas consideraciones que, en mi opinión, permiten llegar a una conclusión general. Voy a intentar demostrar que es probable que no haya ningún nivel semántico al menos en uno de los sentidos tradicionales de esta noción, que no existe un nivel psicológicamente real de representación en el que los términos definibles sean reemplazados por sus definiciones. Si esto es verdad, a fortiori las corrientes generativa e interpretativa dentro de la semántica están equivocadas; el vocabulario primitivo del sistema representacional interno es comparable en riqueza al vocabulario superficial de un lenguaje natural.

Suele aceptarse que el inglés contiene la transformación *equi-NP deletion* (de aquí en adelante = *equi*), o supresión por equivalencia de SN_s, que suprime el sujeto de una oración subordinada en condiciones dc identidad con un *SN* de la oración inme-

diatamente subordinante. La existencia de pares como (8) parece sugerir dicha regla, y esto se ve confirmado por la observación de que se considera que «John objects to being bitten» tiene a «John» como sujeto lógico de *ambos* verbos (es decir, que la oración dice que John se opone a que *John* sea mordido).

8) $John_1$ objects to his_1 being bitten.
[= John se resiste a ser mordido]
John objects to being bitten.

Hasta ahora todo va bien. Sin embargo, lo que queremos destacar es que *equi* tropieza con dificultades evidentes cuando opera en el ámbito de cuantificadores como «only» [= sólo]. Consideremos el caso (9).

9) Only Churchill remembers giving the speech about blood, sweat, toil and tears[24].

[= Sólo Churchill recuerda haber pronunciado el discurso de la sangre, sudor, fatiga y lágrimas)

Supongo que (9) es verdad si a) fue únicamente Churchill quien pronunció el discurso y b) Churchill recuerda que lo pronunció. Si (9) es verdad en estas condiciones, también debe serlo la oración de la que se deriva (9) por la transformación, *equi*[25]. Pero ¿cuál puede ser esa oración? A primera vista, hay tres posibilidades: (10-12).

(10) Only Churchill remembers himself giving the speech...

(11) Only Curchill remembers his giving the speech...

(12) Only Churchill remembers Churchill('s) giving the speech....

Pero, a primera vista, no nos servirá ninguna de estas posibilidades. La oración (10) queda eliminada porque, aunque es equivalente a (9), es de suponer que ella misma es una oración transformacionalmente derivada, y las únicas fuentes disponibles son (11) y (12); por eso, suponer que (9) procede de (10) es sencillamente reemplazar la pregunta «¿De dónde procede (9)?» por la pregunta «¿De dónde procede (10)?». Pero es inmediatamente evidente que (11) y (12) quedan excluidas también, pues ninguna de ellas es equivalente a (9). Por ejemplo, del hecho de que sólo fuera Churchill quien pronunciara el discurso y de que Churchill lo recuerde no se deduce que sea sólo Churchill quien recuerde que pronunció el discurso. Lo que se demuestra con ello es que *yo* recuerdo que él pronunció el discurso, y lo mismo se puede decir, indudablemente, de muchos otros. De igual manera, de las premisas mencionadas no se deduce que sólo sea Churchill quien recuerda que Churchill pronunció el discurso,

[24] De aquí en adelante, abreviado a «Only Churchill remembers giving the speech...». El ejemplo surgió en una conversación con la profesora Judith Jarvis Thomson, a quien expreso mi agradecimiento.
[25] Estoy suponiendo que las transformaciones son «conservadoras del significado», independientemente de lo que *esto* signifique exactamente.

pues, volvemos a repetirlo, *yo* recuerdo que Churchill lo pronunció; el mismo argumento que impide deducir (9) de (11) impide deducirlo de (12). Podríamos decir (en broma) que el recordar haber pronunciado un discurso tiene una forma curiosa de particularidad epistémica: es algo que sólo puede hacer quien pronunció el discurso. Pero recordar que él pronunció el discurso (o que fue Churchill quien lo pronunció) es algo que puede hacer cualquiera que tuviera oportunidad de escucharlo. Da la impresión de que (9) no puede proceder de ninguna de las oraciones (10-12).

Una solución de estos datos requeriría: *a*) salvar la transformación *equi* (es decir, demostrar que (9) no es un contraejemplo); *b*) brindar una fuente para (9) que no sea una fuente posible de (11) y (12); *c*) explicar la relación entre (9) y (10) (es decir, por qué son equivalentes). Quiero considerar a continuación tres soluciones diferentes. Lo que distingue unas de otras es fundamentalmente las suposiciones que hacen (o, en cualquier caso, toleran) sobre el carácter del vocabulario de las representaciones más profundas a que se aplican las transformaciones. Las tres soluciones son compatibles con los datos propuestos hasta ahora, pero veremos, sin embargo, que hay motivos plausibles para elegir entre ellas.

Solucióu 1: Descomposicióu de «ouly» [= sólo]

Comenzamos con una línea de análisis que propone que el cuantificador superficial «only» [= sólo] no ocurre en el vocabulario de los niveles más profundos de la representación lingüística. En concreto, según este análisis, a) «only» no ocurre en el nivel más profundo de representación para el que se definen las transformaciones; b) «only» se introduce en las estructuras superficiales mediante una transformación por lexicalización; c) la lexicalización tiene el resultado de derivar secuencias superficiales de la forma «sólo *a* es *F*» a partir de las secuencias subyacentes de forma aproximada a «*a* es *F* y ningún otro *x* es *F*». Esto constituye, de hecho, un típico análisis semántico generativo, aunque, por lo que yo sé —y por razones que se comprenderán en seguida—, ningún semántico generativo lo ha respaldado.

Vamos a suponer, entonces, dos estructuras básicas: la oración (13) va a ser la representación subyacente a (9) y (10); y (14) va a ser la representación subyacente de (11) y (12).

(13)	Churchill$_1$	remembers	he$_1$	give	the	speech	and
		recuerda	él	pronunciar	el	discurso	y

(no	other$_x$)	x	remembers	(x give	the	speech)
ningún	otro		recuerda	pronunciar	el	discurso

(14) Curchill$_1$ remembers he$_1$ give the speech and (no other$_x$)
(x remembers $\left\{ \begin{array}{l} he \\ Churchill_1 \end{array} \right\}$ give the speech)

Lo que interesa tener en cuenta es que tanto *equi* como la *reflexivización* implican la identidad entre los *SNs* en que actúan, y aunque esta condición se cumple en los

items en cursiva de (13) (es decir las variables), *no* se cumple en los items en cursiva de (14).

Dadas estas estructuras, las derivaciones son rutinarias. O *equi* o la *reflexivización* se pueden aplicar a (13), dando lugar, respectivamente, a (15) y (16)[26].

(15) Churchill$_1$ remembers he$_1$ give the speech and
(no other *x*) (*x* remembers (give the speech))

16) Churchill$_1$ remembers he$_1$ give the speech and
(no other *x*) (*x* remembers (himself give the speech))

Sin embargo, como acabamos de indicar, ninguna de estas transformaciones se aplica a (14), por lo que el *SN* subordinado debe permanecer o como pronombre o como nombre.

Lo que sí se aplica ahora, tanto a (14) como a (15) y (16), es la *reducción conjuntiva,* transformación que suprime opcionalmente en las oraciones coordinadas el primer elemento de un par de *SVs* idénticos. La aplicación de la *reduccion conjuntiva* a (15) y (16) da lugar, respectivamente, a (17) y (18)[27].

(17) Churchill and no other *x* (*x* remembers giving the speech)

(18) Churchill and no other *x* (*x* remembers himself giving the speech)

La aplicación de la *reducción conjuntiva* a (14) da lugar a (19). Téngase en cuenta que ni *equi* ni la *reflexivización* se pueden aplicar a «$\begin{Bmatrix} \text{he}_1, \\ \text{Churchill}_1, \end{Bmatrix}$» en (19) porque el *SN* subordinante (derivado) no es «Churchill» sino «Churchill and no other *x*»; es decir, las condiciones de identidad de *equi* y de la *reflexivización* no se cumplen en (19) más que en (14).

(19) Churchill$_1$ and no other *x* (*x* remembers $\begin{Bmatrix} \text{he}_1 \\ \text{Churchill}^1 \end{Bmatrix}$ give the speech)

Ahora la *lexicalización* se aplica para introduicr «only» en las tres estructuras. La oración (17) se convierte en (9), la (18) se convierte en (10) —derivando así (9) y (10) de la misma fuente última, es decir (13)— y las dos versiones de (19) se convierten en (11) y (12), respectivamente[28].

---

[26] No me voy a ocupar, aquí ni en ninguna otra parte, de las modificaciones morfológicas exigidas por el tiempo, concordancia, etc.

[27] El lector advertido quizá se haya dado cuenta de que, estrictamente hablando, no *tenemos* una identidad entre los *SV* sobre los que va a operar, en (15) y (16), la *reducción de conjunción:* en el primer caso suprime «remembers he$_1$ give the speech» en presencia de «remembers give the speech», y en el segundo caso, suprime «remembers he$_1$ give the speech» en presencia de «remembers himself give the speech». Lo que tiene que ocurrir, de hecho, es que *equi* (en un caso) y la *reflexivización* (en el otro) deben operar sobre *SN* en el conjunto de la izquierda, dando lugar, respectivamente a «Churchill remembers giving...» y «Churchill remembers himself giving...». La *reducción de conjunción* puede actuar ahora en condiciones de identidad estricta con los *SV* derivados para dar lugar a (17) y (18).

[28] Si la *lexicalización* tiene como consecuencia una nueva redacción de «Churchill and no other *x*» en forma de «only Churchill», tendrá el efecto secundario de dejar una variable no ligada en la estructura re-

¿Puede ser cierta esta versión? Por ahora quiero dejar abierta la cuestión de si «only» se reduce a primitivos en un cierto nivel de representación lingüística *más* abstracto que la estructura profunda sintáctica. Lo que parece claro, en cualquier caso, es que *si* se da ese caso, no puede volver a ensamblarse mediante ninguna operación *transformacional.* En concreto, «only *a* is *F*» [= sólo *a* es *F*»] no puede ser una forma lexicalizada de «*a* and no other *x* is *F*» [= «*a* y ningún otro *x* es *F*»] si la lexicalización es un proceso sintáctico. La cuestión es que aceptar este tratamiento equivaldría probablemente a abandonar constricciones ampliamente reconocidas de las transformaciones, y nadie está dispuesto a hacerlo.

El SNC, por ejemplo, dice que las transformaciones (y por lo tanto, la *lexicalización* en especial) deben operar sobre constituyentes. El problema es que la transformación que necesitamos en el caso que nos ocupa debe ser tal que convierta «and no other *x*» en «only» dentro de la frase «Churchill and no other *x*». Pero parece más probable que la distribución por grupos de esa frase sea (20) y no (21).

(20) ((Churchill) (and) (no other *x*))

(21)* ((Churchill) (and no other *x*))

Si esto fuera cierto, una transformación que actuara tal como debería hacerlo la *lexicalización* se aplicaría ipso facto a un no-constituyente. La conclusión tendría que ser que no existe tal transformación.

He dado por supuesto que (21) es un análisis erróneo de «Churchill and no other *x*», pero es posible que haya quienes aceptarían (21) antes que abandonar la descomposición sintáctica de «only». De hecho, aunque por otros motivos, Ross (1967) ha defendido esa distribución parentética. Por ello, puede valer la pena que señalemos que la posible violación de SNC no es la única objeción que se pueda presentar al análisis propuesto. Por ejemplo, «only» es un determinante en «only Churchill», y «and no other *x*», sea lo que sea, no lo es. Por eso, el tipo de lexicalización implicado tendrá a) que sustituir por «only» lo que es, a primera vista, un no-constituyente, y luego, b) cambiar el etiquetado de este supuesto no-constituyente de forma completamente arbitraria. Parece como si los procesos que serían necesarios para sustituir por «only» la supuesta frase de que procede violara claramente las condiciones normales de las transformaciones. Evidentemente, deberíamos evitar admitir tales procesos si hay alguna posibilidad de hacerlo.

Dos nuevas observaciones antes de considerar algunas soluciones alternativas para los datos propuestos por (9-12). En primer lugar, podríamos preservar la *lexicalización* y el SNC si, en vez de derivar «only» de «and no other *x*» deriváramos «only Churchill» de «Churchill and no other *x*». «Churchill and no other *x*» es un componente de (17-19), por lo que una operación que lo sustituye satisface SNC. Pero esto no sirve de mucho. Las dificultades de esta propuesta son precisamente las mismas que hemos mencionado anteriormente al hablar de las definiciones por el uso.

---

sidual. La oración (17), por ejemplo, resultaría «Only Churchill (*x* remembers giving the speech»). Existen formas elegantes de evitarlo, pero no voy a detenerme en ellas, pues, como vamos a ver en seguida, todo el análisis carece de credibilidad.

Hemos supuesto que una de las ventajas de la explicación sintáctica de las definiciones era que nos permitía tener reglas de introducción independientes para lo que son, intuitivamente, items semánticos distintos (por eso, la *lexicalización* introduciría «kill» en las estructuras superficiales y no, por ejemplo, «*x* kill *y*»). Pero ahora, para hacer que la *lexicalización* se conforme a SNC deberemos tener una transformación que introduzca (no «only» sino) «only *a*» en estructuras de la forma «only *a* is *F*». De esta manera vamos contra la intuición de que frases como «only Churchill» no son frases hechas; es decir, que sus significados son constructos formados con los significados de sus términos componentes. Me inclino a pensar que la mayoría de los lingüistas estarían de acuerdo en que sería un precio demasiado alto a pagar por la lexicalización, lo mismo que se admite en general que hay que pagar demasiado por la definición por el uso.

La última observación es que «only» no es el único cuantificador que plantea problemas de lexicalización. Parece que hay un acuerdo filosófico generalizado en el sentido de que «the» [=el, la] se puede definir en función de «*a*» [= un, una] en construcciones como «(el *x* tal que *x* es *f*) es *G*». Más o menos, se supone que la definición es la del número (22).

(22) *Hay una x* tal que (x es F) & (*x* es G) & (*y*)
(*y es F* ≡ (*y* = *x*)).

Como todos sabemos, existe una diversidad de posibles notaciones para expresar esta definición, y existe un notable desacuerdo sobre cuáles de sus cláusulas se presuponen, si es que se presupone alguna, cuando se hace referencia a la *x* tal que *x* es *F*. Parece claro, sin embargo, que todo intento de introducir «the» en las estructuras superficiales mediante un proceso de lexicalización tendría que buscar algún procedimiento para sustituir *(únicamente)* a los items en cursiva de (22). Creo que se puede suponer, sin temor, que cualquier regla que lo consiguiera dejaría, ipso facto, de ser una transformación.

Solucióu 2: Los uombres eu cuauto euautificadores

Da la impresión de que sería conveniente buscar una forma de tratar (9-12) de tal manera que no se presuponga que «only» [= sólo] se descompone sintácticamente. En realidad, ya se han publicado trabajos donde se hace este tipo de tratamiento.

McCawley (1970) ha observado que los pares compuestos de (23) y (24) son aparentes excepciones a la transformación reflexiva.

(23) Only Lyndon pities Lyndon.

[= Sólo Lyndon se compadece de Lyndon].

(24) Only Lyndon pities himself.

[= Sólo Lyndon se compadece de sí mismo].

Por una parte, (23) debería caer en el dominio de la *reflexivización* y, por otra, (23) y (24) no son equivalentes: desde nuestro punto de vista, (24) no es una de las formas superficiales que expresan el mensaje expresado por (23). De hecho, resulta que el que no se pueda hablar de *reflexivización* en (23) es realmente el mismo fenómeno que la imposibilidad de aplicar *equi* en (11) y (12); el fenómeno Churchill y el fenómeno Lyndon son básicamente idénticos, y la solución que sirve para uno de los casos será también válida para el otro.

Pensemos, en concreto, en una solución algo parecida a la que ha propuesto Mc Cawley para (23) y (24)[29]. Supongamos que aceptamos la disponibilidad, en un determinado nivel de representación accesible a las transformaciones, de los mecanismos habituales de referencias cruzadas de la lógica de primer orden. En concreto, presupongamos la distinción entre variables libres y ligadas por una parte y constantes por la otra. Suponemos, además, que el vocabulario de este nivel reconoce no sólo las ligaduras variables clásicas, como *some* [= algo de, algunos] y *all* [= todo, todos], sino también una clase (supuestamente productiva) de cuantificadores «restringidos», que se puede producir de manera uniforme a partir de los nombres propios. Supongamos, en concreto, que «*a*» es el nombre del individuo *a*. En ese caso, el cuantificador correspondiente «(a_x)» es la fórmula tal que «(a_x) $[F_x]$» es verdad si y sólo si todos los miembros de la clase de la que es miembro *a* son *F*. Por consiguiente, podemos definir un cuantificador complejo «(sólo (a_x))», de tal manera que, si «*a*» es el nombre de *a*, entonces «(sólo (a_x)) $[F_x]$» es verdad si y sólo si todos los miembros de la clase de la que *a* es un miembro individual son *F*, y no hay ninguna otra cosa que sea *F*. De esta manera, por ejemplo, si «John» es el nombre de John, en ese caso «(John_x) $[F_x]$» es verdad si y sólo si John es *F*, y «(sólo (John_x)) $[F_x]$» es verdad si y sólo si John es *F*, y ninguna otra cosa es *F*.

Dadas estas convenciones, es posible llegar a un tratamiento razonable del comportamiento de la reflexivización frente a pares como (23) y (24). En concreto, la representación profunda de (23) es algo parecido a (25).

(25) (only (Lyndon_x)) (*x* pities Lyndon)

Aquí no se aplica la *reflexivización*, dado que los dos argumentos de «pities» (es decir, una variable ligada y una constante) no son idénticos. Existe, naturalmente, una fuente de donde procede (24), a saber, (26).

(26) (only (Lyndon_x)) (*x* pities *x*)

En (26) se cumplen las condiciones de identidad en los argumentos de «pities» (ambos argumentos son variables y ambos están ligados por el mismo cuantificador), por lo que se da la reflexivización.

---

[29] Quiero insistir en que la propuesta que estoy a punto de exponer *no* es la que suscribe McCawlay. Los detalles del presente tratamiento están, en realidad, dictados en gran parte por las conveniencias de la exposición. El único aspecto que me interesa, y el aspecto que sí está tomado de McCawley, es la sugerencia de que lo que bloquea la *reflexivización* en (23) y (24) (y bloquearía, *mutatis mutandis*, *equi* en (11) y (12)) es que las condiciones de identidad que imponen estas transformaciones no se cumplen en aquellos pares que se componen de una variable ligada y una constante.

Podría pensarse en la posibilidad de poner pegas a este análisis señalando que (27), igual que (23), no puede ser fuente de reflexivos superficiales.

(27) Only Lyndon pities only Lyndon.

Podría suponerse que (27) cumple las condiciones de identidad de la *reflexivización* y, por lo tanto, según el análisis que estamos realizando, debería dar lugar a (24) como una transformación posible. Esto significaría una complicación, dado que (27) no es más equivalente a (24) de lo que lo es (25). Sin embargo, de hecho el análisis puede resolver las dificultades de (27) tal como ha señalado McCawley; (27) es tratado como un caso de cuantificación múltiple, análogo, por ejemplo, a (28).

(28) Everyone hates everyone.

(= Todo el mundo odia a todo el mundo).

Téngase en cuenta que (28) no puede ser origen del reflexivo superficial (29), y que una forma de solución consistiría en distinguir entre dos cuantificadores en la sintaxis profunda de la misma manera que se haría si estuviéramos «formalizando» (28) en la lógica de primer orden.

(29) Everyone hates himself.

(= Todo el mundo se odia a sí mismo).

En concreto, (28) tendría el análisis profundo (30), donde, como de costumbre, no se aplicaría la *reflexivización* debido a la no identidad de los argumentos del verbo.

(30) (x) (y) $(x$ hates $y)$

Por otra parte, la oración (29) procede de (31), lo mismo que (24) procede de (26).

(31) (x) $(x$ hates $x)$

La simetría es sorprendente y, a primera vista, es un argumento en favor de la propuesta asimilación de los nombres a los cuantificadores.

Lo que queremos insistir en este punto es que (27) no se puede manejar como si fuera fundamentalmente análogo a (28). En concreto, (27) no se reflexiviza si consideramos que contiene dos cuantificadores *diferentes,* y la propuesta presente nos permite hacerlo derivándolo de una fuente como (32).

(32) (only (Lyndon$_x$)) (only (Lyndon$_y$)) $(x$ pities $y)$

Téngase en cuenta que las condiciones de verdad se cumplen perfectamente. Si (27) procede de (32), entonces será verdad si y sólo si todos los miembros del conjunto cuyo único miembro es Lyndon se compadecen de todos los miembros del conjunto cuyo único miembro es Lyndon.

Ahora podemos señalar rápidamente las consecuencias de lo que venimos diciendo para nuestro problema original —qué hacer con (9-12)—. Lo más importante es que exactamente los mismos mecanismos que se utilizaron para impedir que se aplicara la *reflexivización* a (23) se pueden utilizar también para impedir que se aplique *equi* a (11) o (12). De esta manera, (11) y (12) podrían proceder de algo parecido a (33), mientras que (9) y (10) podrían proceder de algo parecido a (34), a través de la *reflexivización* y de *equi*.

(33) (only (Churchill_x)) (x remembers Churchill give the speech)

(34) (only (Churchill_x)) (x remembers x give the speech)

Siguen siendo válidas las consideraciones habituales sobre la identidad: es decir, ni *equi* ni la *reflexivización* se aplican a los pares que constan de una variable ligada y una constante, pero cada uno de ellos puede aplicarse a pares que constan de dos variables ligadas por el mismo cuantificador.

De esta manera tenemos una posible solución de (9-12) que no sólo explica los datos, sino que explica también un número de otros fenómenos sintácticos, aparentemente relacionados. Lo que pagamos por este tratamiento es la postulación de los mecanismos de vinculación de variables en un determinado nivel de análisis *sintáctico* (es decir, en el nivel en el que están definidas las transformaciones). Pero el análisis puede considerarse más bien barato si el precio es ése, ya que es de suponer que tendremos que contar con estos mecanismos *en algún lugar* de la teoría (por ejemplo, en el nivel en que aparecen la ambigüedad de cadenas como «Everybody loves somebody» [= Todo el mundo ama a alguien]). Podríamos concluir provisionalmente que hemos aprendido mucho sobre el vocabulario de las representaciones internas de las oraciones con lo que hemos tratado hasta ahora: que «only» está a nuestra disposición al menos hasta un punto tan alto en las derivaciones como la estructura sintáctica profunda, y que los mecanismos de cuantificación están a nuestra disposición al menos hasta un punto tan «bajo» dentro de las derivaciones como la estructura sintáctica profunda. Sin embargo, creo que las conclusiones basadas en el análisis realizado son prematuras, pues hay razones serias para dudar de que este análisis sea correcto. Vamos a hacer una última prueba con (9-12).

Solucióu 3: *Self* cu la cstructura profuuda

Comenzamos el estudio de los casos de Churchill rechazando la propuesta de que la fuente transformacional de (9) sea (10). Aunque estas oraciones se puedan considerar como equivalentes, los reflexivos suelen tratarse tradicionalmente como formas derivadas[30]. Si el tratamiento tradicional está en lo cierto, y si (9) es el resultado de aplicar *equi* a (10), ¿cuál es la fuente transformacional de (10)?

Pero este argumento únicamente es válido en la misma medida en que lo es la su-

---

[30] Pero ya no ocurre así. La explicación general de la relación entre la *reflexivización* y *equi* que voy a proponer ha sido sugerida independientemente por varios lingüistas, aunque las razones que han presentado no son las que voy a dar yo (cfr. Helke, 1971).

posición de que no hay un elemento reflexivo en la estructura profunda. Supongamos, momentáneamente, que esta suposición es falsa. En concreto, supongamos que *self* es un item del vocabulario básico, y que *equi* se aplica *sólo* a él (es decir, no hay ningún *SN* distinto de *self* que pueda ser suprimido por *equi*). En mi opinión, estas suposiciones resuelven todos los datos que hemos examinado anteriormente: las oraciones (9) y (10) son sinónimas, derivándose la última de la primera mediante la aplicación de *equi;* ni (11) ni (12) pueden constituir la fuente de que procede (9), pues *equi* se aplica únicamente a *self;* ni (11) ni (12) pueden ser origen de (10), ya que no hay ninguna transformación *reflexiva.* Los casos de Lyndon constituyen otro ejemplo de lo que venimos diciendo. En especial, si el reflexivo es una forma básica, no hay ningún problema en bloquear la derivación de (24) a partir de (23). De hecho, según mis conocimientos, si esta proposición es correcta, no hay absolutamente ninguna prueba de la existencia de los mecanismos de cuantificación en ningún nivel específicamente sintáctico de la representación lingüística. Las posibles pruebas se verían refutadas por la no equivalencia de pares como (28) y (29) y la ambigüedad de oraciones como «Everyone loves someone» [= Todos aman a alguien]. Sin embargo, según el punto de vista que estamos adoptando (28) y (29) son formas básicas, y queda sin decidirse la cuestión de si están sintácticamente resueltas las ambigüedades de la cuantificación mixta.

¿Hasta qué punto se puede considerar plausible la afirmación de que *self* es un elemento perteneciente a la estructura profunda? Mi afirmación es que *self* constituye un elemento en el nivel de representación en que se definen la relaciones inferenciales. Para un semántico generativo, con ello se debería dar por zanjada la cuestión pues, por definición, un semántico generativo identifica ese nivel con el más profundo en que se aplican las transformaciones. Sin embargo, si se prefiere la semántica interpretativa, la situación resulta un poco más complicada. Conceptualmente, es posible que *self* aparezca en el nivel semántico, desaparezca en el nivel sintáctico profundo, y luego vuelva a aparecer, introducido transformacionalmente, en oraciones superficiales. Pero aunque, en principio, se trata de una posición posible, no creo que haya nadie que esté dispuesto a adoptarla.

Por eso, lo que quiero hacer ver es que *self* es un elemento en el nivel de representación al que se aplican las reglas de inferencia. Por de pronto, pensemos en el argumento (35). Supongo que este argumento es válido (en términos generales) y que es (aproximadamente) de la forma (36).

(35) a. John believes that Bill is a pothead.
[= John cree que Bill es un pez de colores]

b. Nart believes what John believes.
[= Mary cree lo que cree John].

---

c. Mary believes that Bill is a pothead.

(36) a. John believes S_i

b. $(\exists S_x)$ ((Mary believes S_x & $(S_x = S_i)$)

---

c. Mary believes S_i

Es decir, que el argumento (35) consiste en sustituir el objeto sintáctico de «believes» de la premisa (35a) por el objeto sintáctico de «believes» de la premisa (35b), y la sustitución está autorizada por el hecho de la identidad existente entre lo que cree Mary y lo que cree John. Por el momento, no interesa detenerse en más detalles.

(37) a. The cat wanted to eat the cheese
[= El gato quería comer el queso].

b. The mouse got what the cat wanted.
[= El ratón consiguió lo que quería el gato].

c. The mouse got to eat the cheese.
[= El ratón consiguió comerse el queso].

Doy por supuesto que también este argumento es válido (en términos generales) y que es esencialmente de la misma forma que (35). En especial, doy por supuesto que, tanto en (35) como en (37), la norma relevante de inferencia se aplica a mover el objeto sintáctico del verbo principal de la primera premisa.

Ahora bien, en el caso de (35) estas suposiciones se pueden considerar razonablemente como no problemáticas. Sin embargo, en (37) *existen* ciertos problemas. En concreto, hay que buscar una respuesta a la pregunta: ¿Cuál *es* el objeto sintáctico de «want» en (37a) en el momento en que se aplica la regla de inferencia que autoriza (37)?

Para empezar, hay dos argumentos que permiten pensar que el objeto de «want» debe ser una *oración* en ese nivel. El primero es que, si no es una oración, perdemos la identidad de la forma lógica entre (35) y (37) y esto sería al mismo tiempo antieconómico y contraintuitivo. Nos gustaría aclarar las cosas de manera que (35) y (37) cayeran dentro de la misma regla de inferencia, y, evidentemente, la regla que dirige (35) se aplica a fórmulas con objetos oracionales[31]. En el segundo, parece haber un acuerdo general en que las operaciones inferenciales se definen en relación con objetos tan abstractos *al menos* como las estructuras profundas «standard» (es decir, según Chomsky, 1965). Pero resulta bastante claro que (37a) tiene un objeto oracional en el nivel de estruetura profunda standard. Esto demuestra que las oraciones como (37a) tienen réplicas que contienen complementos pasivizados; cf. (38). Como la *pasivización* se aplica a estructuras de la forma (SN_1 V SV_2), tendremos que suponer que «eat» [= comer]

(38) The cat wanted the cheese to be eaten.

[= El gato quería que fuera comido el queso].

tiene un sujeto *SN* en la fuente sintáctica de (38) y la igualdad del análisis nos exigirá un complemento oracional en la fuente sintáctica de (37a). De ahí es probable que se

[31] Hablando en sentido estricto, el objeto de «believe» en (35) es, probablemente un, *SN* oracional: es decir (creer (que $(S))_{SN}$); la paridad del análisis nos hace pensar en (quiero (que $(S))_{SN}$) para (37). Sin embargo, esto no afecta a la presente argumentación ni en un sentido ni en otro.

deduzca que ambas oraciones tienen complementos oracionales en niveles de representación todavía más abstractos que la estructura profunda standard (por ejemplo, en el nivel semántico) si es que, en realidad, *existen* niveles de representación más abstractos que la estructura profunda standard.

Pero entonces, ¿cuál podría ser el sujeto *SN* incrustado en la representación subyacente de (37a)? Según mi criterio, las opciones notacionales existentes, incluyendo las representadas por las formalizaciones standard de la lógica cuantificacional, se reducen a (39a-c).

(39) a. The cat_1 wented (the cat_1 eat the cheese)

b. The cat_1 wanted (he_1 eat the cheese)

c. (The cat_x) (*x* wanted (*x* eat the cheese))

(39a) corresponde a la suposición de que *equi* se aplica a los *SN* léxicos; (39b) corresponde a la suposición de que *equi* se aplica a los pronombres profundos; (39c) corresponde a la suposición de que *equi* se aplica a variables profundas. En este momento no interesa mucho saber cuál de estas propuestas debe tomarse más en serio, si es que hay alguna que merezca esa consideración. Lo que importa en este contexto es que ninguna de ellas constituye el dominio adecuado de las operaciones inferenciales que hacen posible (37). Dicho a la inversa, los mecanismos disponibles para representar la ligadura y las referencias cruzadas no permiten un tratamiento adecuado de la validez de (37).

Supongamos que (37a) está representado por (39a) en el nivel en que se aplican las reglas de inferencia. En ese caso, la sustitución del objeto sintáctico de «want» [= querer] en la representación subyacente de (37a) en relación con «(what the cat wanted)» [= lo que quería el gato] en la representación subyacente de (37b) dará lugar a la conclusión «the mouse got (the cat eat the cheese)» [el ratón consiguió (el gato comiera el queso)]. Pero es evidente que no es esto lo que dice la conclusión de (37); lo que consiguió el ratón fue (el *ratón* comiera el queso).

(39b) y (39c) no consiguen mejores resultados. Si la conclusión de (37) es que el ratón consiguió que ($él_1$ comiera el queso), entonces o bien (39b) presenta los mismos fallos que (39a) (suponiendo que «he_1» [= $él_1$] constituye una referencia cruzada a «the cat» [= el gato]) o bien «he_1» está funcionando como variable no ligada, y (37c) está representado como oración abierta, cosa que no es, naturalmente. Finalmente, como la subcripción de pronombres y la ligadura convencional de variables son, en este sentido, mecanismos esencialmente idénticos, las consideraciones que excluyen (39b) sirven, *mutatis mutandis,* para excluir también (39c).

El problema está en que, si la regla que da validez a (37) se debe aplicar cambiando el complemento de la representación subyacente de (37a), en ese caso lo que necesitamos como sujeto de ese complemento es, en efecto, no una variable sino una variable de la variable. Es decir, necesitamos una variable que haga una referencia cruzada a «the cat» [= el gato»] en (37a) y a «the mouse» [= el ratón] en (37c). La suposición de que *self* es un elemento del vocabulario de las representaciones a que se aplica la regla, y de que se interpreta como una referencia cruzada al *SN* que lo rige

sintácticamente, nos proporciona precisamente la ayuda que necesitamos[32]. Así, la representación subyacente de (37a) es «the cat wanted (self eat the chese)» [= el gato quería (él mismo comer el queso)] en el nivel en que se definen las operaciones inferenciales. La regla aplicada traslada la oración subordinada a la posición de complemento directo de (37b) dando lugar, como conclusión, a «the mouse got (self eat the cheese)» [= el ratón consiguió (él mismo comer el queso)]. Las convenciones vinculantes de *self* aseguran que constituya una referencia cruzada a «the cat» en la primera oración y a «the mouse» en la segunda, dando lugar precisamente a la representación del argumento que queríamos.

Supongo que estas consideraciones nos inclinan claramente a pensar que *self* es un elemento no analizado en el nivel de la representación semántica; y por lo tanto que es, o es muy probable que sea, un elemento no analizado en el nivel más profundo de la representación sintáctica (dependiendo de si se supone o no que estos niveles son idénticos). Supongo también que de ello se deduce que los mecanismos que utiliza una teoría semántica del inglés para la representación de la referencia cruzada de *SN* son *más ricos* que los mecanismos que utilizan las formulaciones standard de lógica cuantificacional para representar la referencia cruzada de las variables.

Podemos hacer ahora un resumen de la discusión principal. Hemos visto que las pruebas examinadas no exigen una descomposición sintáctica de «only» y que, probablemente, está excluida debido al conflicto con el SNC y otras constricciones que operan sobre las transformaciones. También hemos visto que el tratamiento no sintáctico (el tratamiento que presupone que «only» es primitivo en el nivel en que se aplican las transformaciones) explica los datos analizados, con tal que se suponga que los mecanismos standard de ligazón de variables son posibles en ese nivel. Sin embargo, los principales argumentos en favor de la existencia de estos mecanismos en la estructura profunda dependen de la interacción de los mismos con la supuesta transformación reflexiva, y la prueba de (37) hace que sea plausible pensar que el morfema reflexivo, después de todo, no está introducido transformacionalmente. Considerando todo esto simultáneamente, las conclusiones más adecuadas parecen ser las siguientes:

a) *Self* es un elemento de la estructura profunda; no existe transformación reflexiva.
b) La fuente sintáctica de (9) y (10) es «only Churchill remembers (self give the speech)»; (9) y (10) se diferencian únicamente en que se ha aplicado *equi* en la derivación de la primera.
e) *Equi* sólo se puede aplicar a *self;* en especial, *equi* no se puede aplicar para derivar (9) de (11) o (12).
d) Los pares como (23) y (24) (ó (24) y (29)) no constituyen ninguna prueba en favor de la existencia de cuantificadores y variables en el nivel de la estructura sintáctica profunda. Quizá no existan pruebas de esa índole.

---

[32] Es decir, las condiciones que se han considerado suficientes para que SN_1 reflexivizara a SN_2 en los tratamientos transformacionales de «self», se considerarán ahora suficientes para que SN_2 (= *self*) haga referencia a SN_1; según este tratamiento, el análisis estructural de la supuesta transformación reflexiva se considera como si especificara las condiciones estructurales de la ligazón de *self*.

Me gustaría sacar una moraleja de todo esto. Pero antes de hacerlo conviene observar cierta afinidad espiritual existente entre el fenómeno semántico ilustrado por casos como (37) y el (supuesto) fenómeno sintáctico conocido como «identidad borrosa».

Desde el punto de vista de la notación cuantificacional standard, en que se tienen variables pero no variables de las variables, lo que parece que ocurre en (37) es que las reglas de inferencia son, por así decirlo, «ciegas» a la forma de las variables de los sujetos de las oraciones incrustadas. Es decir, se puede deducir «(ratón$_x$) (*x* consigue (*x* comer el queso))» a partir de la premisa de la forma «(gato$_y$) (*y* quiere (*y* comer el queso))» a pesar de la no identidad de *x* e *y*. Ahora bien, se ha señalado muchas veces (véase Ross, 1967) que ciertas transformaciones sintácticas manifiestan una ceguera parecida ante la necesidad de una identidad rigurosa. Por ejemplo, existe la regla de transformación de *do so* [= ...y también...] que, en casos no tendenciosos, deriva oraciones como (40) de oraciones como (41) bajo la condición de que los *SV* de la oraeión fuente sean idénticos.

(40) John ate Cracker Jacks and so did Mary.
[= John comió Cracker Jacks y Mary también].

(41) John ate Cracker Jacks and Mary ate Cracker Jacks.
[= John comió Cracker Jacks y Mary comió Cracker Jacks].

Lo que ahora nos interesa es que da la impresión de que esta condición de estricta identidad se viola en la derivación de oraciones como (42) pues, si se tiene en cuenta el significado, parece que (42) deberá proceder de (43) sin que los SV_s de (43) sean idénticos.

(42) John broke his arm and so did Mary.
[= John se rompió el brazo y Mary también].

(43) John broke his arm and Mary broke her arm.
[= John se rompió el brazo y Mary se rompió el brazo].

Estos casos indican que *do so* es ciego también a la forma de las variables.

Ahora bien, parece claro que (37) no depende de una identidad borrosa, ya que sus premisas no llegan a estar relacionadas sintácticamente. Por eso, o hay fenómenos paralelos y distintos que explican (37), por una parte, y (42), por la otra, o habrá que reducir el tratamiento de (42) al tratamiento de (37). Este último procedimiento me parece preferible aunque no es, que yo sepa, obligatorio. Es decir, podríamos suponer que (42) procede no de (43) sino de (44), aplicándose *do so* en condiciones de identidad rigurosa mientras que los dos *selfs* se interpretarían mediante la forma de principios de referencia cruzada mencionados más arriba.

(44) John broke self's arm and Mary broke self's arm.

[= John rompió el brazo de sí mismo y Mary rompió el brazo de sí misma].

Esto implica suponer que *self* + *posesivo* + *género* tiene la realización superficial «his/her» [= su, de él/su, de ella], pero esa suposición es plausible de forma independiente: No hay una forma superficial «$\left\{\begin{matrix}\text{his}\\ \text{her}\end{matrix}\right\}$ self's».

Comenzamos esta exposición suponiendo — como hacen la mayoría de los actuales lingüistas, generativistas e interpretativistas — que existe un nivel de representación en el que las palabras quedan reemplazadas por su frases definidoras. Nuestra intención era considerar varios de los procedimientos posibles para conseguir esta sustitución utilizando «only» como medio de comprobación. Desde este punto de vista, los resultados de la investigación parecen poco favorables. No sólo no encontramos ningún procedimiento claramente aceptable para eliminar «only», sino que terminamos por defender una solución que reconoce «only» en el nivel más profundo en que se pueden aplicar las transformaciones, y que reconocer un sistema de referencias cruzadas que es más rico que el empleado por la lógica cuantifieaeional standard en el nivel en el que se define la inferencia.

Se trata, por supuesto, de un ejemplo elegido deliberadamente para crear dificultades, y, naturalmente, es absurdo intentar hacer generalizaciones partiendo de un caso aislado. Pero lo que sí se puede decir cuando menos es lo siguiente: no existe nada en los datos que hemos considerado hasta ahora que nos haga pensar que el vocabulario primitivo de los niveles superiores de la representación lingüística sea notablemente menos rico que el vocabulario superficial del inglés. En especial, ninguno de estos datos nos indica que la sustitución del *definiendum* por el *definiens* sea un proceso significativo en la decodificación de formas ondulatorias en mensajes. De hecho, creo que existen varias razones de peso para mantener cierto escepticismo ante la existencia de semejante proceso, independientemente de las inferencias que pudiéramos sentirnos inclinados a extraer del caso «only». Voy a hacer un comentario breve sobre las mismas.

Lo primero que hay que decir es que toda teoría que mantenga que la comprensión de una oración implica sustituir sus términos definidos por sus expresiones definidoras parece requerir que la complejidad definicional del vocabulario de una oración prevea la dificultad relativa de comprender la oración. En ese caso, la representación canónica de una oración que contenga *W* debe ser más complicada que la representación canónica de una oración que contenga *W'*, dado que *W'* es un primitivo en función del cual se define *W* y suponiendo que todos los demás factores permanecen constantes. (Así, por ejemplo, «John is unmarried» [= John no está cansado] debería ser una oración más sencilla que «John is a bachelor» [= John es soltero] si partimos de la suposición de que «bachelor» [= soltero] se define como «unmarried man» [= hombre no casado]. La representación semántica de «John is unmarried» es *John is unmarried;* pero la representación semántica de «John is a bachelor» es *John is unmarried and John is a man* [= John no está casado y John es un hombre]). Probablemente, esta forma de asimetría debería manifestarse en resultados psicológicos mensurables, pues, si partimos de los supuestos señalados, la representación semántica de la oración *W* puede necesitar más pasos para su computación y, sin duda ninguna, requerirá mayor espacio en la memoria que la representación semántica de la oración *W'*.

Pero, de hecho, parece que no se da la prevista correspondencia entre complejidad definicional y perceptual[33]. En realidad, como ha señalado el Dr. Michael Treisman (en una conversación), si *hubiera* una correspondencia semejante sería difícil comprender cómo la definición abreviatoria explícita puede tener el valor heurístico que de hecho tiene para facilitar el razonamiento. Las abreviaturas (y, por lo mismo, los sistemas de recodificación en general; véase Miller, Galanter y Pribram, 1960; Norman, 1969; Paivio, 1971) no serían de mucha utilidad si la comprensión de una fórmula exigiera sustituir sus términos definidos por las expresiones complejas que los definen. Por el contrario, si la abreviación facilita la comprensión, parece que se debe precisamente a que somos capaces de entender oraciones que contengan las abreviaciones *sin* realizar estas sustituciones.

Convendría recalcar, a la luz de todo esto, que la objeción que estamos proponiendo es igualmente válida en relación con las explicaciones generativa e interpretativa de las representaciones semánticas. El tema en discusión entre estas escuelas se refiere (fundamentalmente) a los mecanismos por los que las definiciones reemplazan a los definibles en el nivel semántico; según la explicación generativa (pero no según la interpretativa), estos mecanismos constituyen casos especiales de transformaciones sintácticas. Lo que interesa ahora, sin embargo, es que no hay razones claras que acrediten la realidad psicológica de ningún nivel de representación en que se hayan definido las expresiones definibles. Volveremos en seguida a la cuestión de cómo se podría elaborar una teoría semántica que no presuponga que la definición es una relación semántica fundamental; por lo tanto, una teoría que postule representaciones internas cuyo vocabulario sea comparable en riqueza al de las oraciones superficiales de un lenguaje natural.

También el punto de que nos vamos a ocupar ahora va dirigido al mismo tiempo

---

[33] Los únicos casos que conozco en los que se han presentado pruebas en favor de esta correspondencia implican fenómenos muy especiales como la marcación lingüística. Así, Clark y Chase (1972) han demostrado que las oraciones que contienen el miembro marcado de un par de palabras suelen ser más difíciles de dominar que las oraciones correspondientes que contienen el miembro no marcado del par (por ejemplo, una oración que contenga $\left\{\begin{matrix}\text{lejos}\\ \text{alto}\end{matrix}\right\}$ es más fácil, *ceteris paribus*, que su oración de control que contiene $\left\{\begin{matrix}\text{cerca}\\ \text{bajo}\end{matrix}\right\}$. Clark y Chase quieren explicar esta asimetría argumentando que la forma no marcada se analiza semánticamente como (negativa + forma no-marcada); por ejemplo, «cerca» = «no-lejos$_{\text{no-marcado}}$». Según este análisis, la diferencia observada en la facilidad de procesamiento podría ser un caso especial de la supuesta correspondencia general entre complejidad psicológica y complejidad definicional.

Aun cuando Clark y Chase tengan razón en esto, es dudoso que se pueda inferir mucho a partir de un fenómeno tan local como la marcación. Pero, de hecho, parece poco probable que Clark y Chase estén en lo cierto, pues parece poco probable que se pueda mantener su análisis de la marcación. Los problemas son muy complicados y no me voy a detener en ellos. Pero, dicho en términos generales, si «bajo» = «negativo+alto$_{\text{no-marcado}}$», donde «alto$_{\text{no-marcado}}$» es el nombre de la dimensión de la altura, en ese caso «John es bajo» es analiza como «es falso que John tenga altura», lo cual, evidentemente, no es así. Esto nos permite pensar que debemos admitir *tres* términos para cada relación de marcación: por ejemplo, «bajo$_{\text{marcado}}$» como en «John es bajo», «alto$_{\text{no-marcado}}$» como en «¿Cómo de alto es John?» y «alto$_{\text{marcado}}$» como en «John es alto». Según esta explicación, «John es alto» y «John es bajo» deberían ser *equivalentes* en complejidad definicional. Por eso, las asimetrías computacionales que presentan no se pueden explicar recurriendo a la complejidad definicional.

contra las concepciones generativa e interpretativa de la semántica. Es el siguiente: ambas clases de teoría postulan una distinción de clase, no justificada, entre fórmulas verdaderas en virtud de las definiciones y otras determinadas clases de «analiticidad».

Las verdades definicionales son, por su propia naturaleza, simétricas. Si «bachelor» significa «unmarried man», en ese caso «unmarried man» significa «bachelor», y de ahí se deduce que «*x* is an unmarried man» implica «*x* is bachelor» si y sólo si «*x* is a bachelor» implica «*x* is an unmarried man». Ahora bien, parece que hay ciertas relaciones semánticas que son exactamente iguales que la que se da entre «bachelor» y «unmarried man», con la excepción de que *no* son simétricas, y la teoría definicional de la analiticidad carece sencillamente de recursos para representar este hecho. El caso clásico es la relación entre, por ejemplo, «red» [= rojo] y «colored» [= de color]. Si es una verdad lingüística que los solteros no están casados, podría considerarse igualmente candidato a la analiticidad el que el rojo es un color. Pero los dos casos se diferencian de la siguiente manera. Es plausible decir que «bachelor» implica «unmarried» porque «bachelor» significa «unmarried man» y «unmarried man» implica «unmarried». Pero no hay ningún predicado *P* tal que sea plausible decir que «red» implica «colores» porque «red» significa «a color and *P*» [= un color y *P*]. Me refiero no solamente a que *no existe* en inglés semejante predicado, sino a que *no puede haberlo* en ninguna lengua; dicho predicado no podría tener ningún significado coherente. Pensemos, por ejemplo, en que tiene sentido perfecto hablar de *x*s que son igual que los solteros con la única diferencia de que no están necesariamente sin casar. Sería una forma circunlocutoria de referirse a los hombres. Pero no tiene sentido que trate de hablar de *x*s que son igual que el rojo con la única diferencia de que no son necesariamente colores. ¿Qué *serían* entonces?

La noción de que las verdades lingüísticas proceden de definiciones requiere que siempre que *Fx* implica analíticamente *Gx* y no viceversa, haya siempre un *H* tal que *G* y *H* sean lógicamente independientes y tal que *Gx y Hx* implique *Fx*. Pero parece que esto no es verdad. El resultado es que las teorías definicionales de la analiticidad o bien ignoran los casos contrarios (igual que han sido ignorados generalmente por los semánticos generativos) o los tratan por medios esencialmente ad hoc (como en Katz, 1972)[34]. Una forma de expresar esto es afirmar que una teoría semántica debe representar la relación entre «bachelor» y «unmarried man» como la réplica bidireccional de la relación unidireccional entre «red» y «colored». Pero ni la corriente generativa ni la interpretativa tienen los recursos para ello. De hecho, ninguna de estas teorías proporciona razones de principio para afirmar que las dos relaciones tengan algo en común.

---

[34] Sospecho que esta clase de casos se extiende mucho más allá de los términos referentes a las sensaciones. (De hecho, lo que *sospecho* es que incluye prácticamente todo el vocabulario no lógico, no sintáctico.) En general, resulta considerablemente más fácil afirmar las condiciones lógicamente necesarias de las expresiones del lenguaje natural que definirlas. Hemos observado más arriba que «kill» no significa, lógicamente, hacer morir, aunque, muy probablemente, es analíticamente imposible matar a alguien sin causar su muerte. Creo que habría que tomar estos hechos con toda seriedad: los mejores ejemplos de las verdades lingüísticas suelen ser asimétricos, que es precisamente lo que no prevé la explicación definicional de la analiticidad. (Para una exposición más amplia, véase J. D. Fodor, a punto de publicarse.)

Si las aplicaciones que se derivan de los términos del vocabulario «no lógico» de un lenguaje natural no dependen de un proceso de definición, ¿cómo se determinan? Una propuesta habitual (desde Carnap, 1956) ha sido la de que si queremos que F implique G (donde uno de los dos o ambos son expresiones morfológicamente simples del lenguaje objeto) deberíamos *decir* sencillamente que F implica G; es decir, deberíamos añadir $\ulcorner F \rightarrow G \urcorner$ a las reglas de inferencia. Estas reglas no estandarizadas de inferencia han recibido el nombre de «postulados de significado», por lo que la propuesta a que nos referimos viene a decir que son los postulados de significado los que realizan la tarea que habitualmente se consideraba realizada por las definiciones[35].

No quiero desarrollar esta propuesta en profundidad: están apareciendo numerosas publicaciones sobre el posible papel de los postulados de significado en el análisis semántico de los lenguajes naturales, y aquí nos limitaremos a remitir al lector a ellas. (Véase, especialmente, Fillmore, 1971; Lakoff, 1970b; Fodor, Fodor y Garrett, de próxima aparición). Bastará con señalar tres de las ventajas más notables.

1. El tratamiento basado en los postulados de significado no exige que la teoría establezca una clara distinción entre el vocabulario lógico y no lógico del lenguaje objeto; la conducta lógica de «bachelor», según este punto de vista, no recibe un tratamiento fundamentalmente diferente de la conducta lógica de «and». Ambos ocurren en el vocabulario del metalenguaje, y las implicaciones que engendran están determinadas por las reglas de inferencia que las rigen.

2. A diferencia de las teorías basadas en la definición, el planteamiento del postulado del significado *no* prevé una correspondencia entre la complejidad de una oración y la complejidad de las definiciones de las palabras que contiene. «John is a bachelor» y «John is unmarried» pueden manifestar las relaciones de complejidad que se prefieran, pues *tanto* «bachelor» *como* «unmarried» ocurren en el vocabulario del nivel de representación en que se especifican los mensajes. En realidad, las reglas de inferencia que gobiernan la relación entre fórmulas en ese nivel determinan que la primera oración implica la segunda; pero la *aplicación* de estas reglas no forma parte de la *comprensión* de la oración (como, según la semántica generativa y la interpretativa, se supone que ocurre con la recuperación de la representación semántica de la oración).

Convendría tener presente que la comprensión de una oración supone computar una representación de la oración que *determina* sus implicaciones; no presupone computar sus implicaciones. (Eso sería imposible; hay demasiadas.) Pero la representación de «John is a bachelor» sí que determina la implicación «John is inmarried» si a) la representación de «John is a bachelor» es ***John is a bachelor*** y b) las reglas de inferencias que se aplican a esa representación incluyen *bachelor→unmarried.*

Estamos suponiendo, en efecto, que el vocabulario superficial de un lenguaje natural es idéntico, o en cualquier caso no mucho mayor que el vocabulario en que se

[35] Desde un punto de vista formal, los postulados de significado podrían parecer precisamente como definiciones por el uso: es decir, podrían aplicarse a expresiones sometidas a análisis sintáctico y en el contexto de variables. Como los postulados de significado no pretenden *definir* las expresiones a las que se aplican, el admitir que una expresión compleja cae dentro del dominio de un postualdo de singificado *no* equivale a afirmar que esa expresión no tiene estructura semántica interna. De esta manera, los postulados de significado nos permiten utilizar los mecanismos formales de definición por el uso sin suscitar las objeciones mencionadas anteriormente.

formulan los mensajes. Como para que lleguen a entenderse las oraciones lo que hay que representar son mensajes, no debe producir extrañeza que no haya ninguna covariación entre las demandas computacionales que impone la comprensión de una oración y la complejidad de las definiciones de las palabras que contiene la oración. El aprender una definición implica principalmente aprender un postulado de significado. Lo cual supone una constricción (no para la memoria computacional de trabajo sino) para la memoria a largo plazo; se añade una regla de inferencia a la lista que se almacena en ésta. Esta es la razón por la que, según la opinión que estamos exponiendo, la definición abreviatoria y otros sistemas de registro facilitan la comprensión de las fórmulas: la memoria computacional es cara, pero la memoria a largo plazo es barata.

Creo que este punto es lo suficientemente importante como para merecer algo más de atención. Parece razonable suponer que una teoría del oyente debe contener dos componentes diferenciados. El primero de ellos se ocupa de explicar la comprensión de las oraciones en sentido riguroso; es decir, de describir las computaciones que realizan la correspondencia entre formas ondulatorias y mensajes; es decir, de determinar las operaciones mentales que terminan en una manifestación de la información que comunican las elocuciones de las oraciones; es decir, de demostrar cómo reconstruyen los oyentes las intenciones comunicativas de los hablantes. Podríamos llamar a este componente «entendedor de oraciones». El segundo componente se ocupa de representar los procesos de datos (incluyendo la extracción de inferencias) que se definen en relación con la información que transmiten las emisiones de oraciones; es decir, los procesos de datos que median la utilización por el oyente de la información que obtiene a partir de las elocuciones que oye. Este componente podría recibir el nombre de *lógica*. De esta manera, y en términos aproximados (haciendo abstracción del feedback y aspectos semejantes), el output del «entendedor de oraciones» es el input de la «lógica». De la misma manera, la (o una) función del entendedor de oraciones es representar las elocuciones en la forma normal con respecto a la cual se definen las operaciones de la lógica.

Ahora bien, dadas las idealizaciones habituales, las operaciones del entendedor de oraciones son operaciones «on-line». Comprendemos una elocución cuando la oímos. Pero las operaciones de la lógica pueden exigir cierta cantidad de tiempo. A veces hacen falta minutos, o días, o semanas para comprender algunas de las implicaciones de lo que hemos oído. Y como por lo general hay un número infinito de implicaciones, podemos tener la seguridad de que hay algunas implicaciones que nunca llegaremos a advertir.

Lo que nos interesa dejar claro es que alguien tiene que cargar con el mochuelo. Supongamos que admitimos que la relación entre formas ondulatorias y mensajes es muy abstracta. Supongamos, en concreto, que admitimos que la sustitución del *definiendum* por el *definiens* se produce *en el proceso de atribuir un mensaje a una forma ondulatoria.* Lo que conseguimos con esta suposición es la correspondiente sim-

---

[36] Por cierto, me parece que una objeción definitiva a los modelos de «redes» del oyente la constituye el hecho de que ni establecen ni admiten esta distinción entre la comprensión de una instancia de oración y el reconocimiento de lo que ella implica. Véase, por ejemplo, Collins y Quillian (1969) y sus herederos espirituales.

plificación de la lógica; ésta ya no necesita tener reglas que especifiquen la conducta del *definiendum* pues, por hipótesis, el *definiendum* ha sido definido antes de que llegáramos a una representación a la que se aplica la lógica. Pero tenemos que pagar un precio para ello: cuanto más sencilla es la lógica, más complicados tendrán que ser los procesos que atribuyen mensajes a las formas ondulatorias.

En resumen, tenemos dos opciones teóricas generales: podemos admitir definiciones en vez de postulados de significado y de esta manera simplificar la lógica al precio de complicar el entendedor de oraciones, o podemos admitir postulados de significado en vez de definiciones y de esta manera simplificar el entendedor de oraciones al precio de complicar la lógica. Lo que quiero dejar claro en este momento es que, *ceteris paribus,* sería más conveniente optar por la segunda posibilidad. Lo importante de la comprensión de oraciones es que es *rápida;* demasiado rápida, en realidad, para que pueda ser explicada por ninguna de las teorías psicolingüísticas conocidas hasta la fecha[37]. Y vamos complicando este misterio en la misma proporción en que convirtamos en abstracta la relación entre formas ondulatorias y mensajes, pues es esta relación la que se pide al entendedor de oraciones que compute. Y a la inversa, reducimos el misterio en la medida en que suponemos una teoría «suave» de los mensajes, pues cuanto mayor sea la semejanza estructural entre lo que se llega a emitir y su representación interna, menor será el trabajo computacional al que tendrá que realizar el entendedor de oraciones. El interés de los postulados de significado está en que proporcionan un procedimiento general para complicar la lógica en formas que reducen el esfuerzo de la comprensión de oraciones. Es decir, nos permiten hacer lo que tienen que hacer las teorías psicológicas: simplificar la representación de las computaciones que hay que realizar «on-line».

3. Según esta explicación, no hay ninguna razón para suponer que la analiticidad dependa de relaciones simétricas. Ciertas reglas de inferencia van en una sola dirección, y otras reglas de inferencia van en doble dirección. No hay nada de especial en esto último.

Quisiera terminar esta sección haciendo desaparecer ciertas incompatibilidades aparentes entre lo que acabo de decir y algunas de las cosas que dije al final del Capítulo 2.

En el Capítulo 2 afirmaba que el lenguaje interno debe ser capaz de expresar la extensión de cualquier predicado que se pueda aprender: es decir, que por cada uno de estos predicados debe haber un predicado coextensivo del lenguaje interno. Pero *no* pretendía demostrar que los niños nacieran con conceptos como el de «aeroplano» ya formados. Por el contrario, afirmaba, lo que deben tener de modo innato son los elementos en que se descomponen estos conceptos, junto con las operaciones combinatorias apropiadas definidas en relación con los elementos. En efecto, es posible reducir los compromisos nativistas de la explicación del lenguaje interno si se presupone que la definición está entre los procesos que se desarrollan en el aprendizaje

[37] Puede verse un cálculo de esta velocidad en la obra de Marslen-Wilson (1973) sobre las influencias semánticas en las tareas de seguimiento («Shadowing»). Estos estudios indican que al menos *cierta* información sobre el contenido del material lingüístico se puede conseguir en el plazo de un cuarto de segundo de su recepción.

de conceptos. Hasta ahora todo va bien. Pero en las páginas anteriores he tratado de reivindicar que, probablemente, los predicados del lenguaje natural no están representados internamente por sus definiciones: la representación del mensaje de «bachelor» es *bachelor* y no *unmarried man.* ¿Cómo es posible compaginar estas afirmaciones?

Creo que debe pensarse seriamente en la siguiente posibilidad: *bachelor* entra en el lenguaje interno *en cuanto abreviación de una expresión compleja del lenguaje interno:* es decir, en cuanto abreviación de *unmarried man.* La convención abreviatoria se almacena como principio de la lógica (es decir, como *bachelor* ⇉ *unmarried man*). Como en el curso del aprendizaje del inglés, «bachelor» se vincula con *bachelor* y «unmarried man» se vincula con *unmarried man, bachelor* ⇉ *unmarried man* se puede utilizar para mediar relaciones inferenciales como la que existe entre «*x* is a bachelor» y «*x* is an unmarried man».

Quiero subrayar que, aunque es posible que esto no sea cierto, no es una tontería. Más bien al contrario, hace posible un número de predicciones empíricas claras. Con este modelo, es de esperar a) que no haya correlación entre la complejidad definicional relativa de un término y la dificultad de comprensión de una oración que contenga dicho término (véase más arriba); pero que b), en ciertos casos haya correspondencia entre la complejidad definicional relativa de un par de términos y el orden en que se aprenden. Como estamos partiendo de la suposición de que el proceso de definición es, por así decirlo, ontogenéticamente real, es de esperar que el niño domine los términos que corresponden al *definiens* antes de que domine los términos que corresponden al *definiendum.* Si, por ejemplo, *only* se define en función de *all,* sería de esperar que «all» [= todo] se aprenda antes que «only» [= sólo]. Y así ocurre en realidad.

Podría presentarse como argumento la posibilidad de demostrar con pruebas empíricas que esta predicción es falsa en sentido general. Así, Brown (1970) ha señalado que la clase de nombres que antes suele aprender el niño son de una abstracción media; «perro», por ejemplo, entra en el vocabulario antes que «animal» o «caniche». Y como es de suponer que «perro» se define en función de «animal», parece que el modelo ontogenético observado por Brown sería incompatible con la teoría que acabo de defender.

Existen, sin embargo, varios problemas en esta línea argumental. En primer lugar, aunque los niños utilicen «perro» antes de utilizar «animal», no se puede descartar completamente que lo que ellos *quieren decir* cuando dicen «perro» es aproximadamente lo mismo que *nosotros* queremos decir cuando decimos «animal», por lo cual las observaciones que nos ocupan no demuestran que el significado de «perro» se consiga antes que el significado de «animal». Es cierto que los niños utilizan inicialmente términos de clase con una generalización excesiva, desde el punto de vista del adulto. Lo que no parece cumplirse es la afirmación de Vygotsky de que el consenso *extensional* medie la comunicación entre niños y adultos.

En segundo lugar, toda nuestra exposición se ha basado en la suposición de que lo que se aprende cuando se aprende un término como «perro» (o «aeroplano», u otros términos de clase semejantes) se representa adecuadamente como un conjunto de condiciones lógicamente necesarias y suficientes. Pero eso, tal como señalé en el Capítulo 2, parece ser algo muy dudoso. Parece suficientemente plausible considerar

que gran parte del conocimiento conceptual se organiza en torno a estereotipos, ejemplares, imágenes, o lo que ustedes quieran, y no, al menos en primer lugar, en torno a las definiciones[38]. (Los problemas revisten en este terreno una enorme dificultad: ¿cómo se tiene *acceso,* por ejemplo, a un ejemplar? Si el concepto de perro es, en gran parte, una representación de un perro estereotípico, ¿cómo se realiza la tarea de determinar qué es lo *cae dentro* del concepto?). Sin embargo, parece que la orientación general va por el buen camino. Lo que media el primer uso de «aeroplano» por el niño no es, sin duda ninguna, el conocimiento de que los aeroplanos son máquinas voladoras. Más bien, las cosas son aeroplanos en cuanto que son como otras cosas que el niño ha visto rondar por el aire. La teoría definicional de los conceptos tiene muy poca cuenta del papel que juega la ostensión en la fijación de los conocimientos.

En resumen, puede ser cierto, como he insinuado anteriormente, que, *en la medida en que un concepto está representado internamente como una definición,* el orden de la adquisición de los términos es paralelo al orden de complejidad definicional de los conceptos que expresan los términos. Pero no podremos comprobar esta afirmación mientras no sepamos cuáles son los conceptos (si es que los hay) que se representan internamente como definiciones, y la información de que disponemos en la actualidad nos hace pensar que muchos de ellos no están así representados.

He aquí un resumen de nuestras conclusiones:

1. Las pruebas lingüísticas examinadas son compatibles con el punto de vista de que el vocabulario de los mensajes (y, a fortiori, el vocabulario de las representaciones internas en general) es muy rico.
2. Si esto es verdad, entonces los procesos de datos que operan sobre los mensajes (es decir, la lógica) deben ser complicados en la misma proporción. Debe haber algo que determine las relaciones conceptuales existentes entre términos «no lógicos» del vocabulario del lenguaje natural, y si no lo hace el entendedor de oraciones, tendrá que hacerlo la lógica.
3. Los postulados de significado son candidatos plausibles para el enriquecimiento de la lógica.
4. Por eso, provisionalmente, la relación entre el *definiendum* del lenguaje natural y el *definiens* del lenguaje natural se expresa mediante postulados de significado definidos en relación con sus respectivas traducciones al lenguaje interno.
5. En concreto, la sustitución de los definibles por sus definiciones *no* es uno de los procesos que media la comprensión de una oración; es característico que *definiens* y *definiendum* tengan representaciones en distintos niveles de mensaje.
6. La disputa entre semántica generativa y semántica interpretativa, en la medida en que es una disputa sobre el tratamiento sintáctico de las definiciones, es una tormenta en un vaso de agua. En el sentido de «definición» de que se

---

[38] Véase Helder (1971), Putnam (sin publicar) y Paivio (1971). Estos tres teóricos, por lo demás diferentes, están de acuerdo en la inadecuación de las definiciones para expresar lo que sabemos sobre las clases.

trata, la definición no constituye una idea central dentro de la teoría semántica.

7. En concreto, no existe ningún nivel de representación (incluyendo el nivel semántico) en que «kill» [= morir] y «cause to die» [= hacer morir], «only» [= sólo] y «none but» [= nadie más que], etc., reciban representaciones idénticas.
8. Estos puntos de vista son por lo general compatibles con las consideraciones relacionadas con la velocidad de comprensión de la frase. Como el procesamiento de las oraciones es muy rápido, son preferibles las teorías que afirmen que la representación de una oración que debe ser recuperada para su comprensión está relacionada en forma relativamente *poco* abstracta con la forma superficial de la oración. Dichas teorías colocan la carga computacional donde más fácilmente se acomoda: en procesos «*off*-line».

La intención principal de este capítulo ha sido fundamentalmente la de ilustrar algunas formas de argumentación que parten de datos sobre los lenguajes naturales para apoyar hipótesis relacionadas con las representaciones internas. El planteamiento, en líneas generales, ha consistido en suponer que algunas representaciones internas representan oraciones, por lo que si sabemos cómo se representan las oraciones sabemos cómo son algunas de las representaciones internas.

Nuestra conclusión es que, muy probablemente, gran parte de la complejidad léxica de las oraciones superficiales puede darse también en el nivel de representación en que se explicitan los mensajes. Esto puede parecer un punto de vista que estaría dentro de una exposición que acepta la metodología de la gramática generativa. Por eso, antes de terminar deseamos hacer una observación metodológica.

Los teóricos —tanto filosóficos como lingüistas— que han considerado seriamente la posibilidad de formalizar los lenguajes naturales han hecho generalmente dos suposiciones sobre el sistema de representaciones que estaban tratando de construir. En comparación con los lenguajes naturales, el sistema representacional se supone que es al mismo tiempo explícito y simple.

Supongo que el requisito de la explicitud *es* precisamente el requisito de la formalidad. Las propiedades semánticas de las oraciones del lenguaje objeto deben ser definibles literalmente en relación con sus traducciones en el sistema representacional. Las reglas para manipular la información transmitida por las oraciones deben aplicarse mecánicamente a las representaciones semánticas que reciben las oraciones. Por otra parte, la simplificidad constriñe la *base* del sistema representacional más que las relaciones entre sus fórmulas y las reglas dentro de las que caen. Un sistema simple (al menos en uno de los sentidos importantes del término) es aquel que tiene un vocabulario primitivo relativamente pequeño y una sintaxis relativamente poco complicada.

Lo que interesa ahora es que, estrictamente hablando, la satisfacción del objetivo de la explicitud no está relacionada conceptualmente con la satisfacción del objetivo de la simplicidad. Esta última implica algo que no implica la primera: que los recursos comunicativos de un lenguaje natural puedan, en principio, ser captados por un sistema que sea estructuralmente menos complicado que los lenguajes naturales. La

suposición de que el inglés puede ser formalizado en algún sistema representacional no exige, sin embargo, que pueda ser formalizado en un sistema cuya sintaxis y vocabulario sean notablemente diferentes del vocabulario y sintaxis superficiales *del inglés*.

Evidentemente, este tipo de consideración debe ser tomado con seriedad por quien desee descubrir representaciones semánticas que sean psicológicamente reales. En definitiva, existen constricciones de las representaciones internas distintas de la maximización de la simplicidad de la base del formalismo en que se formulan; la más importante es maximizar la eficiencia computacional de los procesos de datos definidos sobre dichas representaciones. Los filósofos han solido afirmar no sólo que las oraciones de un lenguaje natural tienen una forma lógica determinada, sino también que su forma lógica se pueda expresar en un sistema más bien parecido a la lógica cuantificacional de primer orden. Los lingüistas han afirmado generalmente no sólo que las reglas semánticas se pueden definir en relación con las estructuras básicas, sino también que el vocabulario y sintaxis de las estructuras básicas es fundamentalmente más sencillo que el vocabulario y la sintaxis de las cadenas superficiales. Lo importante es que las suposiciones formalista y reduccionista podrían, al menos en principio, darse por separado. Si los argumentos que hemos estado considerando son correctos, quizá lo más conveniente sea precisamente separarlas.

Por eso, puede tener cierto sentido la insistencia del Wittgenstein de los últimos años en la riqueza superficial de los lenguajes naturales; en cualquier caso, nadie tiene derecho a dar por descontado que su complejidad es *meramente* superficial en el sentido de que podríamos comunicarnos igual —o mejor— con sistemas formalmente más sencillos. Naturalmente, esto se aplica en ambas direcciones. Si no se puede suponer que un lenguaje adecuado para las representaciones semánticas debe ser menos complejo que los lenguajes naturales, tampoco se puede argumentar contra la posibilidad de formalizar los lenguajes naturales basándose en que son muy complicados. Si las oraciones son objetos complejos, esto sólo puede demostrar que necesitamos un metalenguaje igualmente complicado para representar su forma lógica. En resumen, si la independencia de la reducción y la formalización no siempre ha quedado clara para los formalistas, tampoco ha quedado siempre clara para sus críticos.

El resultado de estas observaciones es una sugerencia que yo considero totalmente especulativa pero muy interesante: a saber, que el lenguaje del pensamiento puede ser muy parecido a un lenguaje natural. Es posible que los recursos del código interior estén representados en forma más bien directa en los recursos de los códigos que utilizamos para la comunicación. Lo menos que podemos decir en favor de esta hipótesis es que, si es verdadera, constituye un paso importante hacia la explicación de por qué los lenguajes naturales son tan fáciles de aprender y por qué es tan fácil entender las oraciones: los lenguajes que podemos aprender no son muy diferentes del lenguaje que sabemos innatamente, y las oraciones que conseguimos entender no son muy diferentes de las fórmulas que las representan internamente.

Parece conveniente terminar insistiendo en que estos puntos de vista pueden estar todos ellos equivocados: incluso en el caso de que se pueda mantener la orientación general del presente capítulo. La tesis que más interés tengo en mantener es que las afirmaciones (o, cuando menos, algunas de las afirmaciones) sobre el carácter de las representaciones son empíricas en el sentido de que los datos empíricos tenderían ha-

cia su confirmación o su rechazo. He tratado de demostrarlo argumentando que los datos sobre los lenguajes naturales se refieren directamente a, y suelen elegir entre, hipótesis competitivas sobre el vocabulario de las representaciones internas que el hablante/oyente atribuye a las elocuciones de oraciones. Lo que interesa dejar bien sentado es que no es necesario que estos argumentos sean decisivos para que la demostración consiga resultados positivos. Lo único que hace falta es que sean argumentos. Es perfectamente posible que las soluciones que he propuesto para los ejemplos examinados resulten inadecuadas. Pero, en ese caso, la prueba deberá referirse a otros ejemplos o a mejores soluciones. En cualquier caso, presupondrá que las teorías sobre la forma y contenido de las representaciones internas deben competir en adecuación metodológica y en adecuación a los hechos, lo mismo que ocurre con otras clases de teorías científicas. Este es, muy en resumen, el contenido del presente capítulo.

Capítulo 4

LA ESTRUCTURA DEL CODIGO INTERNO: ALGUNAS PRUEBAS PSICOLOGICAS

E pluribus unum.

Si es verdad gran parte de lo que he afirmado en los anteriores capítulos, la relación causal entre estímulo y respuesta estará *típicamente* mediada por la representación interna que los organismos se forman de cada uno de ellos. Y si *eso* es cierto, casi todos los resultados de la psicología —desde la psicofísica a la psicometría— se pueden poner en relación, de una u otra manera, con hipótesis sobre la forma de ser del sistema de representaciones internas. Nos encontramos así ante una situación epistémica que es normal para una ciencia viva: en principio, los datos no determinan de forma suficiente las teorías; de hecho, tenemos más datos de los que sabemos utilizar —muchos más de los que pueden manejar nuestras teorías.

Naturalmente, no tengo intención de examinar toda la psicología como forma de demostrar esta afirmación. Lo que haré es concentrarme en una de las conclusiones que parecen deducirse de los trabajos experimentales. Además, me limitaré en general a mi campo. Muchos de los resultados que vamos a examinar proceden de la investigación de los procesos psicolingüísticos. Creo que es muy probable que estos resultados se puedan generalizar a otras áreas de la psicología, pero considero que se trata de una cuestión empírica abierta. Dentro de los objetivos de esta obra, será suficiente si consigo demostrar que hay al menos algunas clases de resultados psicológicos que constriñen la teoría de las representaciones internas que median al menos algunos procesos mentales.

La afirmación que trataré de demostrar es la siguiente: probablemente, es un error hablar de *el* sistema de representaciones internas que el organismo tiene a su disposición para el análisis de los hechos del entorno o las opciones de conducta. Más bien, en circunstancias normales, los organismos tienen acceso a una variedad de tipos y niveles de representación, y el que —o los que— se atribuya en el curso de una determinada computación estará determinado por una serie de variables, incluyendo factores de motivación y atención y el carácter general de la apreciación que hace el organismo de las características de demanda de su tarea. Si la conclusión del

Capítulo 2 fue la riqueza del sistema representacional que debe estar en la base de la percepción y de la integración de la conducta, la conclusión del presente capítulo será la flexibilidad de dicho sistema y la racionalidad de los mecanismos con los que se explota.

Comencemos examinando algunos aspectos de las oraciones y el reconocimiento de las oraciones mencionados en el Capítulo 2. En él señalamos que uno de los principios fundamentales de la lingüística moderna es que toda oración de un lenguaje natural admite un análisis en cada uno de un número determinado de niveles descriptivos. Cada uno de estos niveles tiene las propiedades de un lenguaje formal: tiene su propio vocabulario y sintaxis, y existe una clase propia de entidades abstractas que son los «designata» de sus términos en las interpretaciones que se intenta hacer de los mismos.

Si la descripción estructural atribuida por una determinada gramática a una determinada oración es correcta, las propiedades que señala deben ser precisamente aquellas en virtud de las cuales las elocuciones de la oración se conforman a las convenciones del lenguaje que describe la gramática. En concreto, lo que comunican por norma general las elocuciones de la oración viene determinado por a) cuáles son las convenciones del lenguaje, y b) cuál es la descripción estructural de la oración. Por eso, es razonable suponer a priori que la comprensión de instancias de oraciones implica probablemente atribuirles instancias de descripciones estructurales. Y, como hemos señalado también en el Capítulo 2, contamos ahora con gran cantidad de pruebas a posteriori que nos hacen pensar que esta suposición es verdadera. Como se puede decir lo mismo, *mutatis mutandis,* de la *producción* de oraciones, estamos en condiciones de proponer una primera aproximación a una teoría de los procesos psicolingüísticos: el reconocimiento perceptual de una elocución implica atribuirle una serie de representaciones cada vez más «abstractas» (una para cada nivel de descripción lingüística reconocido por la gramática de ese lenguaje), y la producción de una elocución implica la representación de la conducta que se pretende en cuanto que satisface la correspondiente serie de representaciones cada vez *menos* abstractas, cuyo último miembro se puede interpretar directamente como una matriz fonética[1].

---

[1] Estoy presuponiendo que los parámetros de una matriz fonética determinan el conjunto de *inputs* al aparato vocal en la medida en que el *output* del aparato vocal se puede interpretar como habla (es decir, en la medida en que es interpretable fonéticamente). De la misma manera, un determinado conjunto de valores simultáneos de estos parámetros (en cuanto especificados por la representación de rasgos distintivos de un sonido del habla) corresponde a un estado determinado de excitación de los articuladores (aunque las pruebas existentes indican que esto ocurre sólo de forma muy indirecta —a través de una serie de transformaciones subfonéticas de los valores de la matriz; para más detalles, véase Liberman, Cooper, Shankweiler y Studdert-Kennedy, 1967). El resultado de estas suposiciones es la posibilidad de contar con un esbozo general de lo que sería una respuesta a la pregunta: «¿Cómo se traducen en *conducta* las *intenciones conductuales* en el curso de la producción del habla; en especial, cómo consigue el hablante producir elocuciones que se acomoden de hecho a las descripciones fonéticas a las que *trata* que se acomoden?».

La respuesta sugerida es que cuando las intenciones conductuales son conductualmente eficaces es debido a que a) una de las descripciones a que trata de acomodarse la conducta es interpretable en cuanto conjunto de instrucciones para los órganos efectores pertinentes, y b) la organización fisiológica del sistema es tal que, suponiendo que todas las demás cosas permanecen iguales, el hecho neurológico que codifica las instrucciones excita causalmente a los órganos efectores para que actúen de una manera que sea compatible con las instrucciones (es decir, normalmente, para que las obedezcan). Así, en el caso que nos ocupa,

A estas alturas de la exposición podemos ver ya que «la» representación que se atribuye a una elocución en un intercambio de habla debe ser una especie de objeto muy heterogéneo. Es, efectivamente, la suma lógica de representaciones extraídas de una serie de diferentes sublenguajes del lenguaje interno. Es una cuestión empírica el saber qué tienen en común estos sublenguajes, si es que tienen algo, y algunas de las aportaciones más importantes de la lingüística moderna han sido intentos de responder a esa cuestión (por ejemplo, el descubrimiento de que los niveles morfofonológico y fonético se ocupan del mismo conjunto de rasgos distintivos).

Pero, de hecho, esta explicación es demasiado simple, y las formas en que se aparta de los hechos resultan instructivas. Para empezar con un aspecto trivial, el presente modelo reconoce sólo dos relaciones entre un perceptor y una instancia de oración: o la entiende (en cuyo caso le atribuye una descripción estructural completa) o no la entiende (en cuyo caso no le atribuye ninguna representación en absoluto). Pero es evidente que esto resulta demasiado poco matizado. La comprensión es una noción que admite grados y es posible recuperar una proporción más o menos grande de lo que una determinada elocución trataba de comunicar[2]. Existen varias formas posibles de liberalizar el modelo para conseguir que tenga en cuenta este hecho. Una de las más atractivas es la contenida en las sugerencias hechas por Broedbent (1958).

Vamos a pensar que suponemos que las distintas representaciones lingüísticamente relevantes de una instancia elocutoria se computan literalmente en series de orden creciente de abstracción. Supongamos, también, que, una vez que se ha computado (para cualquier $i > 1$) la representación del nivel i del input, el oyente debe optar por interrumpir la computación o seguir adelante y computar la representación del estímulo en el nivel $i + 1$. De esta manera, cada nivel de representación está asociado a un punto de decisión en que el oyente tiene la posibilidad de no molestarse en seguir computando. Además, en cualquiera de los niveles a) la decisión de continuar con el análisis se debe hacer a la luz de la información sobre el estímulo de que se disponga en ese nivel, y b) la decisión debe hacerse en un tiempo real —probable-

---

una de las descripciones a que trata de acomodarse normalmente la conducta verbal viene dada por una matriz fonética. Pero a') una matriz fonética se puede interpretar como un conjunto de instrucciones al aparato vocal, y b'), si todas las demás cosas se mantienen iguales, el estar en situación de tratar que la propia elocución se acomode a la descripción fonética *D* es causalmente suficiente para excitar al aparato vocal para que produzca una elocución que se acomode a la descripción fonética *D*. (Para que se pueda decir que todas las demás cosas se mantienen iguales hace falta, por ejemplo, que no haya intenciones contrarias y dominantes, que el aparato vocal funcione adecuadamente, y así sucesivamente.) Como hemos indicado previamente, la base en que se apoya la posibilidad de llegar a explicaciones computacionales de la conducta es el hecho (supuesto) de que las relaciones *causales* entre los estados fisiológicos del organismo respetan las relaciones *semánticas* entre las fórmulas del código interno.

[2] Es útil (y probablemente cierto) suponer que una de las cosas que normalmente trata de comunicar una elocución es su propia descripción estructural. (Se trata, naturalmente, de una suposición más fuerte que aquella según la cual una elocución normalmente trata de acomodarse a su descripción estructural.) Cuando hablamos, tratamos de que nuestra elocución se interprete como elocución de una u otra forma verbal, es decir, en cuanto ejemplo de uno u otro tipo lingüístico. Si es cierta la orientación general de la teoría lingüística contemporánea, esta intención se puede identificar con la intención de que el oyente atribuya a la elocución la descripción estructural que individualiza el tipo en cuestión. Lo que nos interesa señalar aquí es que estas intenciones pueden, en un caso determinado, cumplirse a la perfección, o hasta cierto punto, o no cumplirse en absoluto.

mente dentro del tiempo disponible para la exhibición de representaciones del estímulo en la memoria a eorto plazo.

Este tipo de modelo parece intuitivamente plausible, concuerda con el hecho de que hay niveles en la comprensión de una elocución, e incluso existen ciertas pruebas experimentales y anecdóticas en favor de la visión del procesamiento de oraciones que recomienda. El modelo insinúa tres predicciones principales. En primer lugar, si hay realmente «puertas» entre los niveles adyacentes de análisis de tal manera que el input sólo reciba una descripción estructural completa si atraviesa todas esas puertas, habría que esperar que los diferentes estímulos tuvieran diferentes probabilidades de llegar a ser reconocidos y que la probabilidad en el caso de cualquier estímulo determinado estuviera en cierta manera en función de su interés global. En segundo lugar, si las representaciones de los inputs se computan por orden creciente de abstracción, habría que esperar que en el caso de los estímulos que no reciben un análisis completo (por ejemplo, los estímulos a los que sólo se atiende parcialmente) sólo se pudiera hacer mención de la información que es más concreta. Finalmente, como he señalado anteriormente, si la decisión de continuar el análisis se hace en tiempo real, podríamos suponer que la cantidad de representación del nivel i que podría ser pertinente para determinar si se pasa al nivel $i + 1$ debe ser comparable a la cantidad de representación del nivel i que se puede representar simultáneamente en la memoria a corto plazo.

Hay razones para creer que cada una de estas predicciones es verdadera. Las pruebas en favor de la primera son fundamentalmente anecdóticas: parece que todos tenemos la experiencia de que una sensibilidad diferencial a las elocuciones que contienen el propio nombre, o a las elocuciones hechas por una voz conocida, o a las elocuciones que contienen palabras «clave» como «analítico» o «idoneidad». Estas elocuciones parece como si se destacaran del fondo en situaciones de ruido. Según la explicación que estamos exponiendo, esto se debe a que existe literalmente una tendencia a favor de su reconocimiento y del análisis completo de las elocuciones que las contienen. Una fiesta de sociedad (el caso del «cocktail party») parece una especie de experimento natural hecho para confirmar esta afirmación.

En el caso de las otras dos predicciones, podemos recurrir a resultados experimentales bien conocidos. Anne Treisman (1964) hizo una serie de estudios sobre percepción de oraciones en los que utilizó lo que se ha venido a conocer como paradigma de las tareas de seguimiento («shadowing»). En estos estudios el sujeto escucha señales grabadas en una cinta que se presenta dicóticamente a través de auriculares, con una señal diferente en cada lado. *S* recibe instrucciones de atender únicamente a un canal. Sin embargo, al final de la presentación se pregunta a *S* sobre el material presentado por el canal desatendido. El resultado más frecuente es precisamente lo que se podía predecir de lo dicho anteriormente: *S* puede informar únicamente de los rasgos del input desatendido que están determinados en forma relativamente directa por sus propiedades acústicas más aparentes: por ejemplo, que la señal era una voz hablada y cuál era el sexo del hablante, pero *no* el contenido de lo que se decía. Evidentemente, estos resultados son perfectamente compatibles con una visión «de abajo arriba» de la percepción del habla, de tal manera que las representaciones de la señal se computen por orden creciente de abstracción comenzando con la recuperación de sus propiedades acústicas/fonéticas. Parece ser que los mecanismos atencionales

interactúan con las conveniencias para determinar hasta qué punto va a ser completo el análisis realizado de una determinada señal. (En la situación investigada por Treisman, las conveniencias del sujeto están determinadas fundamentalmente por su intención de acatar las instrucciones experimentales de atender únicamente a un canal.)

Una de las variantes del paradigma de Treisman tiene una importancia especial para la tercera de las predicciones enumeradas más arriba. En este caso, el material del canal no atendido es *el mismo* que el material del canal al que debe atender *S*. Sin embargo, la segunda señal va rezagada en relación con la primera por un intervalo que el experimentador puede variar. El resultado es que el reconocimiento por *S* de que los dos canales tienen la misma señal depende críticamente de la magnitud de este intervalo. Dos *S*s no suelen reconocer la identidad de las señales cuando el intervalo es de más de unos 2 segundos y es raro que no la reconozcan cuando el intervalo es menor.

Parece razonable suponer que estos 2 segundos representan el período durante el cual la señal no atendida está disponible en la memoria a corto plazo. Esta interpretación encaja perfectamente con el modelo de Broadbent, que exige algún mecanismo que conserve la información (relativamente) no interpretada durante el tiempo suficiente como para permitir decisiones sobre la conveniencia de realizar un mayor procesamiento. Ampliando la metáfora utilizada más arriba, si la atención es una puerta a través de la cual debe pasar la información de entrada para ser reconocida, los resultados de Treisman permiten pensar que, en el caso del material lingüístico, la puerta se abre durante unos dos segundos. Tiene cierto interés observar que este cálculo de unos dos segundos en cuanto intervalo crítico es al menos ampliamente compatible con las evaluaciones sobre la amplitud de la memoria a corto plazo en relación con materiales lingüísticos hechos con paradigmas experimentales independientes. Véase, por ejemplo, Jarvella (1970), donde se sugiere que el almacenamiento «on-line» del material sintácticamente estructurado contiene unidades de hasta una frase de longitud, aproximadamente, y Crowder y Morton (1969), donde se calcula un plazo de unos 2 segundos para el almacenamiento «ecoico» de los estímulos lingüísticos.

Comenzamos refiriéndonos al hecho de que no todo lo que se oye se entiende plenamente. El modelo Broadbent-Treisman explica el hecho suponiendo que, aunque hay algunos inputs que reciben representaciones en todos los niveles de descripción, hay otros muchos que no. De esta manera, el modelo insiste en el *carácter incompleto* del análisis que reciben algunas elocuciones. Sin embargo, también hemos señalado que existen otras posibilidades de abordar el estudio de estos hechos, y al menos una de ellas la debemos citar aquí.

Para Broadbent y Treisman, existe una puerta entre los niveles adyacentes de descripción y únicamente los inputs que reciben atención plena atraviesan todas las puertas. Los trabajos realizados recientemente por Lackner y Garrett (1973) permiten pensar, por el contrario, que incluso los inputs no atendidos reciben descripciones en los niveles superiores, aunque las representaciones sólo son *accesibles* (por ejemplo, están a disposición del sujeto para que informe de ellas) en el caso de las señales que son objetos de atención.

Igual que los sujetos de Treisman, los de Lackner y Garrett oían materiales lingüísticos en ambos canales de sus auriculares estereofónicos. Y, también como en el

paradigma de seguimiento («shadowing»), la atención de *S* se dirigía a uno de los dos canales. Además, Lackner y Garrett trataron de conseguir la relativa inaccesibilidad del material no atendido descendiendo sustancialmente su volumen en comparación con el del canal atendido. De hecho, el volumen de los canales estaba lo suficientemente desproporcionado como para que, en las entrevistas posteriores a los tests, muchos de los sujetos no fueran capaces de decir siquiera si el canal desatendido contenía habla.

En los dos canales los materiales de estímulo que Lackner y Garrett presentaban a sus sujetos diferían no sólo en volumen sino también en contenido. En concreto, en los casos críticos, el canal atendido contenía una *oración ambigua,* mientras que el canal no atendido contenía un *contexto que eliminaba la ambigüedad.* Por ejemplo, para un determinado sujeto en una prueba concreta, el canal atendido podía contener una oración como (1) mientras que el canal no atendido contenía la (2). La actuación de este sujeto se comparaba con la de sujetos que oían la misma oración en el canal atendido pero cuyo canal no atendido contenía (3) (es decir, un contexto que favorece la otra alternativa en eliminación de la ambigüedad de (1)). Se indicaba a todos los Ss que parafrasearan la oración atendida al final de cada ensayo de tal manera que los experimentadores pudieran determinar qué interpretación le habían dado.

(1) The spy put out the torch as our signal to attack.

El espía {apagó* / sacó} la antorcha como señal de que debíamos atacar.

(2) The spy extinguised the torch in the window.

El espía apagó la antorcha de la ventana.

(3) The spy showed the torch from the window.

El espía mostró la antorcha desde la ventana.

Garrett y Lackner argumentaban de la siguiente manera: si no se analizara ninguna información procedente del canal no atendido, o si sólo se analizara la información de nivel relativamente bajo, entonces el contenido del canal no atendido no podría tener ninguna influencia en el carácter de la paráfrasis que *S* hacía de la oración atendida; de las dos posibles lecturas, las paráfrasis de los *S*s deben reflejar una u otra interpretación aproximadamente en la misma proporción en que lo hacen aquellos sujetos de control en los que el contenido del material no atendido no está relacionado con la interpretación de la oración atendida. Por otra parte si, en relación con el material no atendido, se computan las representaciones de nivel alto, es posible que parte de esta información «consiga entrar» e influir en la paráfrasis de la oración atendida, que de esta manera se vería influida por la dirección dc la señal eliminadora de la ambigüedad. Por sorprendente que parezca, es esta última predicción

* El verbo inglés «to put out» puede significar las dos acciones: apagar y sacar. *(N. del T.).*

la que confirman los datos. Incluso los sujetos que son totalmente incapaces de *informar del* contenido del canal no atendido manifiestan la influencia de su contenido en la elección que hacen de una paráfrasis de la oración atendida. Parece ser que parte de la información sobre el contenido semántico de la oración no atendida se computa aun cuando el sujeto no pueda disponer de ella de forma consciente.

Estos resultados nos trazan una imagen de la relación entre percepción y atención muy diferente de la que proponían Broadbent y Treisman. Si Garrett y Lackner están en lo cierto, la atención funciona no para determinar hasta qué punto recibe el input una representación plena, sino más bien cuál es la parte de la representación de la que se puede informar. Sigue habiendo una «puerta», pero, según la opinión de Garrett y Lackner, está colocada entre la memoria temporal (en la que se computa el análisis estructural del input) y una memoria relativamente permanente en la que los resultados de las computaciones están a disposición de la conciencia. Sólo el material atendido pasa del almacenamiento temporal al permanente, y sólo se puede informar de lo que está en almacenamiento permanente.

Tal como están las cosas, no es fácil determinar cuál de estas explicaciones está en lo cierto, suponiendo que lo esté alguna de ellas. Sin embargo, tampoco interesa para el desarrollo de nuestro tema, pues lo que los datos demuestran incontrovertiblemente es que el modelo de todo o nada (o se cuenta con una representación total del input o nada) no resulta convincente. Si Broadbent y Treisman tienen razón, no siempre computamos el análisis total de lo que oímos. Si la razón está de parte de Garrett y Lackner, gran parte de lo que computamos no llega a almacenarse el tiempo necesario para poder informar de ello. En cualquier caso, parece que el oyente tiene gran libertad para decidir cómo manipular la representación interna de un estímulo. Conviene recordar que, tanto en los estudios de Treisman como en los de Garrett y Lackner, la diferencia entre lo que ocurre a los estímulos rivales está en función de variables *instruccionales;* es decir, las diferencias de procesamiento están determinadas, al menos en parte, por la decisión de *S* de atender al material de un canal y de ignorar el material del otro.

Uno de los temas centrales de este libro ha sido que, para saber qué respuesta va a provocar un determinado estímulo, hay que averiguar qué representación interna atribuye el organismo al estímulo. Evidentemente, el carácter de estas atribuciones debe depender a su vez de cuál es la clase de sistema representacional de que se dispone para mediar los procesos cognitivos del organismo. Sin embargo, ahora lo que nos interesa subrayar es que no depende solamente de eso. Según el modelo Broadbent-Treisman, son los mecanismos atencionales los que determinan cómo se explotan las capacidades representacionales disponibles. En el modelo Garrett-Lackner son los mecanismos —cualquiera que sean— que afectan al paso de la información desde la memoria de trabajo a la memoria a largo plazo. En uno u otro modelo, los estados psicológicos del organismo se ven implicados en la determinación de cuál de las representaciones del estímulo potencialmente disponibles es la que de hecho media la producción de la conducta. Dicho en forma más general, la explotación por el organismo de sus capacidades representacionales es, de forma sistemática, sensible a sus conveniencias. Entre las tareas de una teoría del sistema representacional está la de ayudar a explicar esta interacción.

Pensemos en otro tipo de pruebas para apoyar estas consideraciones. Uno de los

primeros experimentos sobre la realidad psicológica de las gramáticas generativas fue realizado por Mehler (1963). Fodor, Bever y Garrett (1974) explican el tema detalladamente. Aquí nos limitaremos a recordar que Mehler utilizó un paradigma en el que se pedía a los sujetos que memorizaran listas de oraciones de una serie de tipos sintácticos diferentes (por ejemplo, declarativas activas simples, pasivas, negativas, preguntas) y que los resultados indicaban claramente que el tipo sintáctico es uno de los determinantes del nivel de recuerdo. Hablando en términos aproximados, podríamos decir que la probabilidad de que una oración fuera recordada correctamente estaba en relación inversa con la complejidad de su descripción estructural sintáctica, y la probabilidad de que un par de oraciones se confundieran era proporcional a su semejanza sintáctica. (Puede verse un estudio semejante y con resultados comparables en Clifton y Odom, 1966). Por eso, Mehler concluía que la descripción estructural sintáctica de una oración es —o, en cualquier caso, forma parte de— la representación de la oración que llega a almacenarse en la memoria a largo plazo.

Por otra parte, Jacqueline Sachs (1967) presentó a sus sujetos un texto leído, probando el recuerdo de algunas oraciones elegidas al final de cada presentación. Las oraciones estímulo utilizadas cambiaban dentro de las mismas clases de dimensiones sintácticas que las de Mehler, y sin embargo los resultados de su experimento eran claramente distintos. Sachs no apreció prácticamente ninguna influencia de las variables sintácticas; lo único que contaba era el contenido. Es decir, las oraciones sinónimas solían combinarse independientemente de su forma sintáctica, y las oraciones sintácticamente semejantes se distinguían en la medida en que diferían en cuanto a su significado.

¿Qué se puede hacer con esta anomalía? En concreto, si el trabajo de Mehler es un argumento *en favor* de una compenetración específica entre estructura sintáctica y memoria permanente, ¿habría que decir que los resultados de Sachs son contrarios a ella? La respuesta parece ser ésta: el carácter sobresaliente de las variables estructurales depende de la naturaleza de la tarea experimental. En concreto, depende de lo que el sujeto considere que constituye el sentido de realizar la tarea. Wanner (1968) demostró que es posible hacer aparecer y desaparecer el «efecto Mehler» *manteniendo constantes los materiales de estímulo* y según cómo se den las instrucciones al sujeto. Los *S*s a quienes se dice que están participando en un experimento de memoria acusan la influencia del detalle sintáctico; los *S*s a quienes se dice que lean el texto fijándose en el contenido no lo acusan. (Johnson-Laird y Stevenson (1970) hablan de resultados semejantes.) Después de todo, esto no tiene por qué resultar demasiado sorprendente. Sabemos por propia experiencia que es muy distinta la manera de tratar un texto cuando se intenta memorizarlo y cuando se está leyendo sin más. Si se dan instrucciones de que se recuerde, se intenta recordar todo lo que se lee; si las instrucciones son en sentido de que se busque el contenido, se deja de lado todo lo que no sea esencial. Tenemos la sospecha de que la diferencia de tratamiento resulta productiva; que las dos actitudes ante el material producen generalmente representaciones del estímulo almacenadas de distinta manera. Y el estudio de Wanner confirma esta sospecha.

Me parece que todas estas consideraciones apuntan hacia un rasgo fundamental y generalizado de los procesos cognitivos superiores: *el tratamiento inteligente de las representaciones internas*. La psicología seria comienza con el reconocimiento de que

importa la forma en que el organismo especifica los estímulos y las opciones de respuesta. Presupone, por tanto, un lenguaje interno lo suficientemente rico como para representar los inputs que puedan afectar a la conducta y los outputs que pueda desplegar el organismo. Pero ahora parece que existe un ámbito dentro del cual el organismo puede elegir cómo deben explotarse sus recursos representacionales; la conclusión reiterada de los resultados que acabamos de examinar ha sido que el sujeto puede controlar cuáles son las representaciones que se atribuyen a las instancias de oraciones y/o cuál de las representaciones atribuidas se almacena. Al ejercer este control, el sujeto manifiesta una correspondencia racional entre su actuación y (lo que considera que son) las características de demanda de la tarea experimental.

Con esto volvemos a un camino ya muy trillado. Si el sujeto debe elegir entre formas de representar el estímulo y la respuesta, deberá también contar con formas de representar sus opciones; es decir, deberá contar con formas de representar sus maneras de representar el estímulo y la respuesta. Pero tener formas de representar las maneras de representar los inputs y outputs es tener un sistema representacional *en capas*. Algunas expresiones del lenguaje interno se refieren a expresiones *del lenguaje interno*. Y otras expresiones del lenguaje interno se refieren a inputs y outputs (potenciales o reales). Las computaciones cuyas consecuencias determinan la forma en que deben desplegarse los recursos representacionales del sujeto utilizan esencialmente expresiones de la primera clase.

La concepción general de (algunos) procesos mentales superiores implícita en estas observaciones se conoce bastante bien gracias a la obra de los psicólogos cognitivos cuyas especulaciones se han visto influenciadas por la organización de los ordenadores (cf. Miller, Galanter y Pribram, 1960; Newell y Simon, 1972). Uno se imagina una jerarquía de programas «ejecutivos» que funcionan para analizar las macrotareas en microtareas. Estos programas pueden «llamarse» mutuamente y llamar a rutinas de solución de problemas de un nivel inferior, aunque la amplitud de estas referencias cruzadas está limitada por la ingeniosidad del programa y, naturalmente, la capacidad computacional global de la máquina. Cuando las cosas van bien, los resultados de los procesos de orden inferior se pueden integrar para dar lugar a una solución de cualquier macroproblema que se hubiera planteado originalmente al sistema. El que, en un caso determinado, las cosas vayan bien está determinado en parte por el hecho de que los programas ejecutivos consigan seleccionar las subrutinas adecuadas y aplicarlas en el orden debido.

No obstante, lo que nos interesa en este momento no es confirmar el carácter general de esta especie de modelo ni siquiera examinar sus detalles. Se trata más bien de insistir en lo que estas teorías implican en relación con el carácter y reclutamiento del sistema representacional en cuyas fórmulas habría que definir las computaciones postuladas. Las implicaciones más importantes tendrían carácter doble. En primer lugar, como hemos visto, tiene que haber recursos para representar las representaciones. Si una de las funciones ejecutivas es decidir cuáles son las descripciones de nivel inferior que llegan a computarse, el lenguaje que habla el ejecutivo (es decir, el lenguaje en que se definen las computaciones ejecutivas) debe tener procedimientos para referirse a las descripciones que pueden ser asignadas por las rutinas de nivel inferior. En segundo lugar, va implícito en el modelo el hecho de que el carácter de las representaciones desplegadas en cualquier nivel dependerá, en parte, del resultado de

las computaciones de nivel *superior*. En la jerga técnica, el flujo de información en estos sistemas presenta feed*back* de las decisiones de nivel superior así como feed-*forward* de las decisiones de nivel inferior.

Conviene detenerse un momento a reflexionar sobre estos dos puntos. Por una parte, las representaciones internas son lábiles y la eficiencia con que se despliegan puede, en algunos casos, determinar significativamente la eficiencia del procesamiento mental. Por la otra, no conocemos ninguna constricción general de la forma en que fluye la información en el curso de las computaciones que determinan dichos despliegues: cuando decimos que estamos ante un sistema de feedback estamos admitiendo sencillamente que hay otros factores distintos de las propiedades del input que pueden afectar a la representación que recibe el input. En concreto, cuáles son las representaciones internas que se atribuyen es algo que depende en cierta manera del estado cognitivo —según nuestros conocimientos, de *todo* el estado cognitivo— del organismo estimulado. Quizá haya límites a las opciones de que disponen en este sentido los organismos, pero si los hay no hay todavía nadie que sepa dónde se deben colocar. La psicología se las trae.

Pensemos sólo en otro ejemplo más que ilustra la flexibilidad con que se explotan los recursos del sistema de representación interna. Hemos visto que en los modelos standard de solución de problemas se utiliza con frecuencia como una estrategia primaria el análisis de las macrotareas en microtareas. El resultado de esta descomposición de la tarea suele ser el de establecer una jerarquía de objetivos computacionales a largo y a corto plazo, y el flujo de la información dentro de la jerarquía exigirá normalmente que las soluciones de los problemas de nivel inferior hagan de inputs para los procesos de nivel superior. (Véase, por ejemplo, el concepto de unidades-TOTE desarrolado por Miller, Galanter y Pribram, 1960; y Miller y Johnson-Laird, de próxima aparición). Donde se observan estos requisitos estrictamente, toda computación de nivel i debe hacerse antes de que se pueda iniciar cualquier computación del nivel $i + n$. Sin embargo, de hecho, muchas veces es posible avanzar sin cumplir demasiado a la letra dichos requisitos, con tal de estar dispuestos a aceptar incurrir en errores ocasionales. Supongamos, por ejemplo, que los resultados de alguna de las computaciones del nivel i son parcialmente redundantes con los resultados de algunas de las otras. En ese caso podemos *predecir* los resultados de las computaciones últimas basándonos en el hecho de haber *realizado* únicamente las primeras. Como la probabilidad de que la predicción resulte cierta varía en proporción directa con la magnitud de la redundancia, tendremos razones para aceptar la predicción siempre que tengamos razones para suponer que la redundancia es elevada. Evidentemente, podría haber casos en que lo razonable sería aceptar la predicción, pues de esa manera se reduce el número de computaciones que hay que realizar en conjunto.

Dicho en pocas palabras, la carga computacional asociada a la solución de una clase de problemas se puede reducir algunas veces optando por procedimientos de resolución de problemas que sirven sólo *la mayoría* de las veces. En estos casos se da más importancia a la eficiencia que a la fiabilidad. Pero hay procedimientos para apostar sobre seguro. Por regla general, en primer lugar se recurre a los procedimientos heurísticos; cuando fallan éstos, se recurre a procedimientos relativamente más lentos, pero relativamente algorítmicos. Esta forma de ordenar recursos computacio-

nales disponibles puede constituir muchas veces el arreglo ideal entre la velocidad de la computación y la probabilidad de conseguir los resultados adecuados.

Estas son, evidentemente, consideraciones ya sabidas que están en la base de la noción de programación heurística. Nuestra intención en este momento es subrayar que tienen una importancia considerable para las teorías de la representación interna. Como, por norma general, las rutinas heurísticas prescinden de las computaciones que deben realizar los algoritmos, es también frecuente que produzcan análisis de sus inputs relativamente más pobres. Un procedimiento infalible debería representar todas las propiedades de su input que *pudieran ser* relevantes para la tarea. Un procedimiento heurístico puede arreglárselas representando solamente las propiedades de su input que tienen *probabilidad* de ser relevantes para la tarea. Pero, desde el punto de vista que nos ocupa, esto significa que el que un input determinado reciba una descripción determinada en una ocasión determinada es algo que depende, *inter alia,* de cómo se ordenan las conveniencias del organismo: de los pesos relativos atribuidos a la fiabilidad y a la eficiencia al ocuparse de la tarea que se tiene entre manos. Quiero detenerme brevemente en un caso que servirá de ilustración de estos principios.

Hemos visto que un modelo de comprensión de oraciones es, en realidad, un dispositivo que asocia formas ondulatorias con mensajes. Es muy poco lo que se sabe sobre la forma en que podría operar dicho dispositivo, aunque yo me atrevo a pensar que, si comenzáramos ahora y nos pusiéramos a trabajar en ello con ahinco, quizá pudiéramos elaborar uno en quinientos años, más o menos. En cualquier caso, hay una o dos cosas que parecen claras; entre ellas, que todo procedimiento de reconocimiento que pretendiera ser infalible tendría que deducir el mensaje codificado por la instancia de una oración a partir de una especificación de las relaciones gramaticales que se dan entre sus constituyentes. Es decir, si dicho dispositivo debe servir para *todas y cada una* de las oraciones del lenguaje, en ese caso toda subrutina que dé lugar a una representación de un mensaje debe tener, entre sus inputs, una representación de las relaciones gramaticales que se dan en la oración a que se atribuye el mensaje. Partiendo de la suposición cómoda (aunque probablemente falsa) de que un reconocedor de oraciones es un dispositivo totalmente serial, esto se puede convertir en una afirmación sobre el orden de las operaciones en tiempo real: hay que atribuir a una instancia una representación de las relaciones gramaticales *antes de* que se le atribuya la representación de un mensaje.

Creo que se puede considerar como verdad conceptual la afirmación de que, dadas las idealizaciones adecuadas, una persona que hable *L* con fluidez puede ser considerada como un instrumento infalible de reconocimiento de oraciones de *L*. Si, por ejemplo, existen oraciones en inglés que no puede entender ningún hablante inglés, la razón será una limitación de su tiempo, memoria o atención y no, indudablemente, una limitación de su comprensión *del inglés.* En una primera aproximación: Ser una oración inglesa *es* ser algo que los hablantes ingleses *en cuanto* hablantes ingleses pueden entender. Así, si es cierto que los reconocedores infalibles de oraciones *deben* deducir los mensajes a partir de las representaciones de las relaciones gramaticales, parece desprenderse que los hablantes ingleses *pueden* deducir mensajes a partir de las representaciones de las relaciones gramaticales.

Pero aunque es de suponer que puedan hacerlo, se puede constatar que muchas

veces no ocurre así. Lo que parece ocurrir es que las relaciones gramaticales se computan únicamente cuando falla todo lo demás. Existen procedimientos heurísticos de reconocimiento de oraciones que, en efecto, pasan por alto las relaciones gramaticales y deducen los mensajes directamente de su contenido léxico, aceptando, por lo mismo, los riesgos de la falibilidad.

Las oraciones denominadas autoincrustadas («self-embedded») constituyen un caso claro aunque, como veremos, hay otros casos que resultan más interesantes.

En primer lugar, *es* posible aclarar el significado de una oración como:

«*The boy*	*the girl*	*the man*	*knew*	*wanted to marry*	*left in a huff*»*
El chico	la chica	el hombre	conocía	quería casarse con	se marchó enfadado
(2)	(4)	(6)	(5)	(3)	(1)

Lo único que hace falta es tiempo, paciencia y perspicacia para ver que la oración es estructuralmente análoga a, por ejemplo,

«*The girl*	*my friend*	*married*	*makes pots*»
la chica	mi amigo	se casó con	hace macetas
(1)	(3)	(2)	(4)

De hecho, lo que se hace al aclarar estas oraciones es una computación de las relaciones gramaticales existentes entre sus frases, como atestiguan los tipos de errores que se tiene mayor probabilidad de cometer en el proceso. Así, si tenemos problemas con «The boy the girl the man knew wanted to marry left in a huff», lo más probable es que a) intentemos leer «the boy the girl the man» como una frase nominal compuesta (véase Blumenthal, 1966), y/o b) que intentemos considerar «wanted to marry» como complemento objeto de «know» (véase Fodor, Garrett y Bever, 1968). Los más avanzados pueden tratar ahora de interpretar «*Bulldogs bulldogs bulldogs fight fight fight*» como una oración y no, por ejemplo, como un grito de ánimo a un equipo de fútbol. (Pista: interpretar los dos primeros verbos como transitivos [= «atacar a», en vez de «luchar»]).

Lo que nos interesa ahora es que hay un procedimiento más rápido que se puede aplicar a los casos de auto-incrustación. Pensemos en la relativa transparencia de

«*The boat*	*the sailor*	*the dog*	*bit*	*built*	*sank*»
la barca	el marinero	el perro	mordió	construyó	se hundió
(2)	(4)	(5)	(4)	(3)	(1)

Lo que ocurre en este caso es lo siguiente: se considera que la oración es un anagrama, y el mensaje se deduce de consideraciones como éstas: los barcos (pero no los perros ni los marineros) se hunden muchas veces; los marineros (pero no los perros) construyen muchas veces barcos; los perros (pero no los marineros) suelen morder, y cuando lo hacen es más probable que el objeto de su acción sea un marinero y no un

* Los siguientes ejemplos no tienen correspondencia exacta al español. Para que el lector se haga cargo de su complejidad, nos limitaremos a indicar la correspondencia de las palabras y su orden de traducción. Hay que añadir pronombres relativos, que en inglés están omitidos. *[N. del T.]*.

barco. Y así sucesivamente. Parece plausible admitir que en ningún momento se recurre a una descripción estructural sintáctica cuando se trata de reconocer una oración de este tipo. O si se hace, la descripción estructural que se intenta conseguir se deduce probablemente del análisis del mensaje más que al revés. (Pueden verse experimentos relacionados con este tema en Schlesinger, 1968.) Y volvemos a encontrarnos en el punto de partida: si queremos saber cuál es la representación que se atribuye a un determinado input, debemos saber algo sobre las clases de procedimientos computacionales (incluyendo los atajos heurísticos) de que dispone el sujeto para atribuir representaciones a los inputs. Y hay que saber algo sobre cuál de estos procedimientos se ha utilizado de hecho.

Las autoincrustaciones son casos curiosos de la psicolingüística, por lo que conviene señalar que podrían extraerse conclusiones semejantes de otras clases de ejemplos. Considérese el caso de las pasivas. Es un hecho comúnmente admitido (aunque no de forma totalmente unánime) que las oraciones pasivas suelen ser más difíciles de entender que sus correspondientes oraciones activas, y que es posible medir este mayor grado de dificultad. La explicación habitual de este hecho presupone a) que la atribución de relaciones gramaticales a las pasivas precede a la atribución de mensajes, y b) que la atribución de relaciones gramaticales a las pasivas se ve complicada por las propiedades de su forma superficial. En concreto, el *sujeto* superficial de una pasiva es de hecho su *objeto* gramatical, mientras que el verdadero sujeto gramatical aparece como el objeto superficial de una preposición. Hay que desenmarañar toda esta complejidad al mismo tiempo que se realiza la atribución de relaciones gramaticales, y para atribuir mensajes hay que tener distribuidas las relaciones gramaticales. Por eso, las pasivas *deberían* resultar más difíciles de entender que las activas.

Sin embargo, hay que mencionar el hecho curioso de que existen ciertas pruebas de que esta asimetría computacional entre activas y pasivas sólo se da en casos especiales. Por ejemplo, véase Slobin (1966) y la experimentación de Wall presentada en Walker, Gough y Wall (1968). Pero quizá podamos entender la clave de todo ello de esta manera: supongamos que distinguimos entre oraciones «reversibles» e «irreversibles» basándonos en este principio: una oración es reversible si, y sólo si, (o, mejor, en la medida en que) su plausibilidad se mantiene inalterada al cambiar su sujeto gramatical por su objeto gramatical; las oraciones irreversibles son aquellas que no son reversibles. (No hemos hecho un alarde de precisión, pero en el contexto en que nos movemos bastará con lo que hemos dicho.) Así, «Mary was bitten by John» [Mary fue mordida por John] es una pasiva reversible, y «John bit Mary» [John mordió a Mary] es una activa reversible (*véase* «John was bitten by Mary» [John fue mordido por Mary] y «Mary bit John» [Mary mordió a John], ambas correctas). Pero «The ice cream was eaten by the child» [El helado fue comido por el niño] y «The dog bit Mary» [El perro mordió a Mary] son, relativamente, irreversibles (ya que *¿the child was eaten by the ice cream* [el niño fue comido por el helado] y *¿Mary bit the dog* [Mary mordió al perro] son construcciones, cuando menos, dudosas)[3].

---

[3] Debe quedar claro que la reversibilidad no es un fenómeno *sintáctico;* es decir, el que una cadena sea o no reversible *no* es algo que lo determinen sus propiedades formales. La reversibilidad tiene relación con las expectativas de los hablantes sobre lo que es *probable que sea verdad,* y por lo tanto pertenece con todo derecho a la «pragmática».

Los datos de que disponemos nos indican[4] que se da una asimetría computacional entre activa y pasiva únicamente cuando se comparan pasivas *reversibles* con activas *reversibles*. Probablemente, la explicación esté dentro de las orientaciones mencionadas al hablar de las oraciones autoincrustadas. Si se puede deducir el mensaje transmitido haciéndolo directamente a partir del vocabulario de la oración de entrada, se hace así, eliminando de esta manera la necesidad de computar las relaciones gramaticales. Esto es posible en el caso de las irreversibles, por lo que las asimetrías de carga computacional producidas por factores sintácticos suelen desaparecer cuando se trata de estas oraciones. Con las reversibles, sin embargo, no hay posibilidad de recuperar el mensaje que se intenta comunicar, a no ser por el camino más largo; hay que computar el análisis sintáctico que trataba de satisfacer la elocución. Por eso, los rasgos sintácticos predicen la carga computacional cuando las oraciones son reversibles.

Venimos ocupándonos de ciertas pruebas psicológicas que hablan en favor de la proposición de que los procesos cognitivos superiores ponen de manifiesto la distribución inteligente que hace el organismo de sus recursos representacionales. Dentro de unos límites (y por medios) que todavía no hemos llegado a conocer, el organismo es capaz de configurar su asignación de representaciones en forma que reflejan sus expectativas de lo que puede ayudarle a conseguir sus objetivos. Termino este examen señalando que esta capacidad de controlar los recursos representacionales parece tener un interesante desarrollo ontogenético.

Pensemos, una vez más, en la asimetría entre oraciones reversibles e irreversibles. Anteriormente hemos indicado que el oyente puede eludir la computación de relaciones sintácticas en los casos en que el mensaje que intenta comunicar el hablante se puede deducir a partir a) del contenido léxico de su elocución, y b) de la información anterior sobre los mensajes que los hablantes tratarán *probablemente* de comunicar. Es claro que la confianza en estas deducciones dará lugar a veces a error. Pero, por definición, cuanto más irreversible sea una oración, más improbable es que se equivoque la heurística con tal que los hablantes traten generalmente de decir lo que resulta plausible decir. En cualquier caso, la utilización de esta especie de atajo presupone un cierto grado de perfeccionamiento no sólo en relación con los contenidos del léxico, sino también con las intenciones probables de los interlocutores en un intercambio de hablar. Los datos indican que hace falta cierto tiempo para adquirir esta sofisticación y que los niños suelen cometer ciertas clases características de errores hasta que la consiguen.

Bever (1970) presenta los resultados de una serie de estudios del desarrollo de procedimientos heurísticos para el procesamiento de oraciones en el caso de niños de poca edad. Véanse, por ejemplo, los resumidos en la Figura 4-1. Las dos curvas superiores representan, respectivamente, el rendimiento de los niños en el caso de activas totalmente reversibles (por ejemplo, «the cow kisses the horse» [la vaca besa al caballo]); y de activas *i*rreversibles plausibles (por ejemplo, *The mother pats the dog* [La madre acaricia el perro]). La curva inferior representa su rendimiento en lo que podríamos llamar activas irreversibles invertidas (es decir, activas irreversibles con una interpretación *i*mplausible como podría ser «*The dog pats the mother*» [El perro

[4] O, al menos, la mayoría de ellos. Pueden encontrarse pruebas en sentido contrario en Forster y Olbrei (1973).

acaricia a la madre]). La configuración general de los resultados no debe producir extrañeza. El rendimiento de los sujetos es casi perfecto tratándose de irreversibles plausibles, como podría esperarse suponiendo que a los dos años de edad el niño cuenta con los procedimientos básicos para analizar oraciones simples SN V SN. Tampoco es extraño que su rendimiento con las irreversibles invertidas sea relativamente bajo al comienzo y tienda a mejorar con la edad; son éstas precisamente las oraciones en que la heurística basada en las suposiciones sobre las probables intenciones del hablante puedan llevar a error al niño. Por eso, el rendimiento relativamente bajo en el caso de las oraciones poco plausibles representa probablemente el exceso de confianza del niño en estos procedimientos heurísticos, y la tendencia del rendimiento a mejorar con la edad representa probablemente el desarrollo de su conocimiento sobre la forma de hacer sobre seguro sus apuestas heurísticas. Sin embargo, lo que tiene un interés especial es el descenso en el rendimiento con las irreversibles implausibles a la edad de tres años. Se observa que a esa edad los niños obtienen con dichas oraciones *peores* resultados que sus controles de dos años. Bever opina que un rasgo típico del niño de tres años es su gran confianza en las estrategias heurísticas de análisis perceptual; que, en realidad, esta dependencia determina una «etapa» evolutiva que aparece en una gran variedad de tareas experimentales. Si esta explicación es cierta, la extraña desventaja manifestada por los niños de tres años pro-

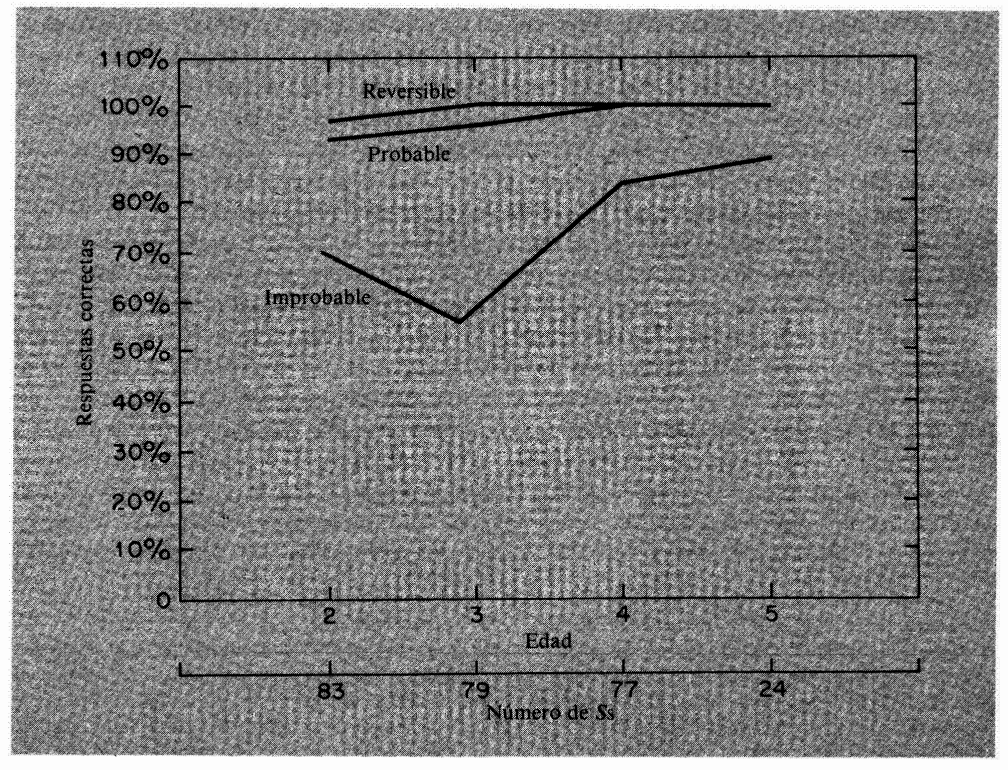


Figura 4-1. Proporción por edades de las respuestas correctas a oraciones activas reversibles, oraciones activas probables, y oraciones activas improbables. Tomado de Bever, 1970, p. 304.

cede del hecho de que han abandonado las rutinas relativamente algorítmicas de procesamiento de oraciones para adoptar procedimientos heurísticos más arriesgados (pero más rápidos). En un momento situado entre los tres y cuatro años comienzan a aprender a controlar su forma de explotar estos procedimientos, a conseguir un equilibrio más realista entre eficiencia y seguridad.

Si las cosas son así, los datos de Bever manifiestan una modulación bastante detallada del rendimiento del niño como resultado del desarrollo de una destreza para controlar sus recursos representacionales. En cualquier caso, el aprendizaje lingüístico del niño debe llegar con el tiempo a una comprensión de las clases de descripciones estructurales que satisfacen las oraciones de su lenguaje; como hemos visto, sólo *porque* satisfacen estas descripciones, las elocuciones de oraciones pueden servir como vehículos convencionales para la expresión de intenciones comunicativas. Pero, por lo que podemos ver, el niño aprende más cosas. Aprende también que, cuando las circunstancias son las indicadas, es posible calcular las intenciones comunicativas a partir de un análisis muy aproximado del carácter lingüístico de la elocución. Y aprende también, dentro de los límites de la falibilidad humana, a distinguir cuáles son las circunstancias adecuadas[5].

Me parece oportuno hacer un intento de recopilación de lo dicho. El objetivo fundamental de este libro ha sido demostrar la existencia de un lenguaje interno en que se realizan las computaciones que están en la base de los procesos cognitivos. En este capítulo, sin embargo, el acento lo hemos puesto no tanto en la existencia de este sistema cuanto en su despliegue. Las principales conclusiones alcanzadas hasta el momento son éstas: en primer lugar, parece existir una variedad de representaciones que puede recibir un input determinado, y el que sea una u otra la representación recibida depende, *inter alia,* de las exigencias de la tarea del sujeto. En segundo lugar, el hecho de que el sujeto consiga emparejar la explotación de sus capacidades representacionales con las exigencias de la situación experimental constituye en cuanto tal una forma de conducta inteligente. No pretendo decir que dichas actuaciones sean conscientes; por el contrario, supongo que generalmente no lo son[6]. Más bien, lo importante es que, cuando las cosas salen bien, lo que el sujeto consigue mediante el dominio de las representaciones internas es una correspondencia racional entre su actuación y sus objetivos. Visto desde una perspectiva opuesta, lo importante es que la representación interna de un estímulo depende no sólo del carácter del estímulo y del carácter del sistema representacional, sino también de las conveniencias del sujeto.

---

[5] Estas observaciones conectan, de forma muy clara, con una larga tradición de trabajos psicológicos sobre los estereotipos, prejuicios y «sesgos perceptuales». Lo que revelan todos estos trabajos es la tendencia del sujeto a «rellenar» los rasgos del precepto que se pueden deducir plausiblemente a partir de (lo que el sujeto considera como) conocimiento previamente adquirido. La conclusión general es la disposición de *S* a conseguir la eficiencia computacional a costa de imprecisiones o falsas representaciones ocasionales. Véanse algunos estudios en Bartlett (1961), Bruner (1957) y Heider (1971).

[6] Como es natural, existen muchos casos de manipulación consciente, voluntaria y estudiada de las representaciones internas en beneficio de una u otra ganancia para la eficiencia computacional. De particular interés es el uso de sistemas mnemotécnicos para facilitar la evocación de materiales estimulares que de otra forma parecerían desordenados; muchos de estos sistemas se basan precisamente en la manipulación disciplinada de las representaciones internas atribuidas a los estímulos. Véase, por ejemplo, los procedimientos mnemotécnicos en que se recurre a hacer versos. (Para un estudio más detallado, cf. Miller, Galanter y Pribram, 1960; Paivio, 1971; Norman, 1969. Luria (1968) estudia algunos casos curiosos.)

Si la línea argumental central de este libro es correcta, el lenguaje del pensamiento constituye el medio para representar internamente los aspectos psicológicamente destacados del entorno del organismo; en la medida en que es posible determinarlo en este lenguaje —y sólo en esa medida— esta información cae dentro de las rutinas computacionales que constituyen el repertorio cognitivo del organismo. Estas rutinas se definen, por así decirlo, sólo en relación con las fórmulas del lenguaje interno. Pero quisiera añadir algo más: algunos organismos, al menos, parecen tener una libertad considerable para determinar cómo se va a utilizar este sistema representacional, y que dicha libertad suele explotarse generalmente de forma racional. En el caso de los seres humanos adultos, al menos, el despliegue de los recursos representacionales parece muchas veces una estrategia calculada para la consecución de objetivos de conducta. Sin embargo, como hemos señalado anteriormente, la existencia de estas estrategias tiene consecuencias importantes relacionadas con el carácter del código en que se realizan. Si los sujetos *calculan* realmente cómo se deben desplegar las representaciones internas, estos cálculos deben definirse también en relación con las representaciones; es decir, en relación con las representaciones de las representaciones. En resumen, ciertas propiedades del lenguaje deben estar representadas en el lenguaje del pensamiento, pues la capacidad de representar representaciones constituye, probablemente, una precondición de la capacidad de manipular representaciones de forma racional.

Estas reflexiones plantean una serie de interrogantes que sería de esperar encuentren respuesta cuando la psicología cognitiva llegue a un grado superior de desarrollo: ¿Cuál es la riqueza de la capacidad del código interno de cara a la auto-representación? ¿En qué medida se explota realmente esta capacidad en la integración de una u otra clase de conducta? ¿En qué medida difieren los individuos en este sentido? ¿Y las especies?

Pero independientemente de las respuestas que lleguen a darse a estas preguntas, hemos avanzado lo suficiente como para comprender que una teoría razonable de la cognición debe distinguirse incluso de los tratamientos más sofisticados asequibles dentro de los límites del asociacionismo. Lo cual puede servir como digno colofón de esta sección.

Podría pensarse que hablar de las representaciones internas no viene a ser, a la larga, algo muy distinto de la adición de un eslabón o dos a las cadenas de estímulo/respuesta. Una opinión semejante ha prevalecido en la psicología «mediacional», que trataba de interponer *representaciones* del estímulo y la respuesta entre los *E*s y las *R*s que reconocen las teorías estrictamente conductistas. (Véase, por ejemplo, Hull, 1943; Osgood, 1957; Berlyne, 1965). Pero con todo y con eso los asociacionistas mediacionales *son* asociacionistas. Como los conductistas más explícitos, postulan vinculaciones mecánicas (o probabilísticas) entre los estados psicológicos y suponen que los vínculos están forjados por las leyes que determinan la fuerza de los hábitos. Las representaciones internas, en concreto, se consideran asociadas a los *E*s y *R*s de la misma manera que están (o se supone que están) asociados entre sí *E*s y *R*s.

Lo que queremos subrayar en este punto es que esta opinión es errónea en todos los sentidos posibles. Las representaciones internas suelen estar emparejadas con lo que representan mediante procesos computacionales (más que asociativos). Es decir,

las representaciones no se *provocan* sino que, por así decirlo, se asignan; y el que sea una u otra la representación asignada está determinado por cálculos que favorecen razonablemente a las necesidades del organismo. Quizá haya —o deba haber— un límite en esta jerarquía de decisiones racionales. Pero no llegamos a ver este límite. Por lo que podemos saber, la cognición está totalmente saturada de racionalidad.

Hasta ahora, la exposición desarrollada en este capítulo se ha centrado en algunos aspectos de lo que se llama algunas veces teoría de la «actuación». Es decir, hemos supuesto que se dispone de un sistema representacional muy poderoso, pero supuestamente monolítico, que es el medio de los procesos cognitivos, y hemos señalado algunas de las opciones que se explotan para determinar cómo se emplea este sistema representacional. Las teorías lingüísticas, y ciertas teorías psicológicas, suelen hacer caso omiso de la existencia de dichas opciones precisamente porque su objetivo es describir en su integridad las capacidades representacionales del organismo. De esta manera, los lingüistas estudian descripciones estructurales completas, aunque llegan a reconocer de buena gana que la computación de descripciones estructurales completas es quizá una estrategia a la que se recurre en última instancia al comprender las oraciones. Los psicolingüistas suelen hacer experimentos en que sólo sirven las estrategias que se pueden considerar como recursos últimos; quizá porque suponen que dichas estrategias son lo que deben tener en común los miembros de una comunidad de habla, mientras que los procedimientos heurísticos pueden variar ampliamente de un sujeto a otro. En cualquier caso, aunque hemos defendido una considerable flexibilidad en las formas en que se *utiliza* el lenguaje del pensamiento, todo lo que hemos dicho hasta ahora es compatible con la opinión de que es *un* lenguaje; que las modalidades de representación interna constituyen, en cierto sentido razonable, un todo uniforme y sistemático.

Existen, sin embargo, razones para dudar de que esto sea cierto. Tradicionalmente se ha afirmado que, junto a los mecanismos representacionales discursivos de que puedan disponer los organismos, existe también una capacidad de representación basada en imágenes o «imaginística» y que la explotación de esta capacidad es fundamental en una serie de funciones cognitivas. Creo que con las pruebas existentes en la actualidad se podría pensar que tal afirmación tiene muchas probabilidades de ser cierta. Por ello conviene decir algo sobre las imágenes incluso en una exposición sumaria de las formas en que los resultados empíricos de la psicología pueden limitar las teorías de las representaciones internas.

Entre los psicólogos que se toman en serio lo de que el pensamiento implica un sistema representacional, la cuestión más tratada ha sido la relación existente entre los items de ese sistema y las cosas representadas por los items; es decir, en términos generales, la cuestión de cómo los pensamientos hacen referencia a los objetos del pensamiento. En este campo existe una doctrina heredada de la tradición empirista británica dentro de la filosofía: los pensamientos son imágenes mentales y se refieren a sus objetos sólo en la medida en que (y sólo en virtud del hecho de que) se les parecen.

Se trata, evidentemente, de una doctrina muy fuerte —mucho más fuerte que la afirmación de que hay imágenes mentales y de que desempeñan un papel ocasional o incluso esencial, en ciertos procesos cognitivos. Insisto en la distinción porque hay

argumentos bastante convincentes en contra de la primera opinión. Si una imagen se parece a aquello a lo que se refiere, el pensamiento no puede ser simplemente cuestión de tener imágenes. Pero sólo conseguiríamos complicar las cosas (lo cual, en cualquier caso, parece una epidemia en este campo) si supusiéramos que como pensar no puede consistir en tener imágenes, de ahí se desprende en algún sentido que no hay imágenes o que, incluso en el caso de que las haya, no pueden desempeñar un papel esencial en el pensamiento. Lo primero que voy a intentar hacer en este apartado es revisar brevemente las consideraciones que demuestran que el pensamiento y las imágenes no pueden ser lo mismo. Luego trataré de considerar la hipótesis más débil, que las imágenes desempeñan un papel interesante en el pensamiento. Acabaré haciendo algunas especulaciones sobre cuál puede ser este papel.

Es probable que ningún psicólogo cognitivo de la actualidad piense que *todos* los pensamientos son imágenes. Hoy en día es más frecuente postular una dimensión de «abstracción» en relación con la cual pueden variar los pensamientos, ocupando las imágenes fundamentalmente el extremo más «concreto»[7]. Ciertos pensamientos concretos son imágenes (según dicen), pero el vehículo del pensamiento abstracto es discursivo.

Para encontrar una versión totalmente desarrollada de la teoría de las imágenes hay que dirigir la mirada a las obras de carácter evolutivo. Bruner, Werner y Piaget (en algunas de sus obras) han propuesto variantes de la opinión según la cual el desarrollo cognitivo del niño está condicionado por una transición que va de las modalidades «imaginísticas» a las modalidades discursivas de representación interna. En líneas muy generales, en el niño de menos edad el vehículo de pensamiento tiene cierta relación no simbólica con sus objetos; los primeros pensamientos *se parecen* a las cosas de las que son pensamientos. Pero el curso del desarrollo está orientado hacia una abstracción creciente en la relación de pensamientos y cosas. Los pensamientos propios de los adultos son (o, en cualquier caso, pueden ser) totalmente simbólicos; es decir, puede haber una semejanza arbitrariamente pequeña entre el vehículo del pensamiento y su objeto; es decir, los pensamientos adultos pueden ser arbitrariamente diferentes de las imágenes.

En la obra de Bruner, por ejemplo, se nos invita a considerar que el niño atraviesa tres etapas evolutivas más o menos distintas, cada una de ellas caracterizada por su modalidad característica de representación interna[8]. En la etapa inicial, el vehículo del pensamiento es un esquema motor interiorizado. (En esto Bruner respalda explí-

---

[7] Es posible que haya imágenes que sean evocadas por términos abstractos. Pero, aunque las haya, no se pueden parecer a lo que denotan dichos términos (por ejemplo, en la forma en que una imagen de John evocada por la emisión de la palabra «John» podría parecerse a John). No hay nada que pueda tener un aspecto semejante a la virtud, por ejemplo, pues la virtud no tiene ningún aspecto. Creo que los argumentos en contra de la identificación de las ideas abstractas con las imágenes son suficientemente conocidos desde Berkeley, aunque todavía se ve de vez en cuando en ciertas obras una tendencia a la confusión en este terreno. Paivio (1971) presenta ejemplos muy instructivos.

[8] Con esto no hacemos demasiada justicia a la sutileza de las ideas de Bruner, pues afirma al mismo tiempo que puede haber un solapamiento en las capacidades representacionales que están disponibles en un determinado momento de la carrera ontogenética del niño, y que entre las diferentes formas de representación pueden existir relaciones de traducción. Sin embargo, lo que a mí me interesa, aquí y en el resto del texto, no es hacer una revisión de todo lo publicado sino sencillamente examinar algunas de las opciones teóricas.

citamente la idea piagetiana de la inteligencia «sensorio-motora».) En la segunda etapa, los pensamientos son imágenes (descritas por Bruner como representaciones organizadas en el espacio más que en el tiempo y que conservan rasgos perceptuales de sus objetos). Finalmente, en el pensamiento maduro, el medio de representación es simbólico en el sentido en que lo son las palabras: no hace falta que haya ninguna semejanza entre el vehículo de la representación y la cosa que representa. Como dice a veces Bruner, en este nivel superior de representación «no se puede saber lo que representa un símbolo por su mera percepción» (1966, p. 31)[9]. Evidentemente, la principal ruptura ontogénica se da entre las etapas dos y tres. En las dos etapas iniciales es la supuesta semejanza entre los pensamientos y sus objetos lo que, dicho crudamente, permite soldar unos con otros. Pero es precisamente la *falta* de esta semejanza lo que constituye la propiedad distintiva de las representaciones de la etapa tercera[10].

Es interesante, para empezar, que este aparato teórico más bien complicado se vea respaldado fundamentalmente por observaciones que son fragmentarias y de carácter impresionista, cualquiera que sea el punto de vista del que las considera. La mejor forma de ilustrar la tenue conexión entre los datos y la teoría es recurrir a una cita. En *Studies in Cognitive Growth* (Estudios sobre el desarrollo cognitivo), Bruner cita observaciones como la del caso de Piaget (1954).

> A la edad de 0:6 Lucienne... coge con la mano el material que cubre los lados [de su palangana]. Tira de los pliegues hacia sí pero se le escapan en cada intento. Entonces se lleva la mano hasta los ojos manteniéndola fuertemente cerrada y la abre con cuidado. Mira atentamente a los dedos y vuelve a comenzar. Esto se repite durante más de diez veces.
>
> Por lo tanto, le es suficiente haber tocado un objeto, creyendo que lo coge, para que considere que está en su mano aunque ya no lo sienta en ella. Este patrón de conducta... manifiesta el grado de permanencia táctil que el niño atribuye a los objetos que ha cogido con la mano (página 22).

No se trata, pensaríamos en seguida, del tipo de datos que podrían contar con un considerable peso teórico. He aquí, sin embargo, cómo los utiliza Bruner:

> Para el niño, por lo tanto, las acciones evocadas por los hechos-estímulo pueden servir en gran parte para «definirlos». A esta edad es incapaz de diferenciar claramente entre percepto y res-

---

[9] Como las imágenes —icónicas o motoras— están especialmente mal adaptadas para ser vehículos del pensamiento abstracto (véase nota 7), el progreso del niño a través de las diferentes etapas constituye también un progreso en dirección a unas capacidades representacionales cada vez más abstractas. En la elaboración teórica de Bruner suelen utilizarse en forma intercambiable «abstracto» y «simbólico».

[10] Bruner, como la mayoría de los escritores que se han ocupado de la naturaleza del simbolismo, presupone que existe una distinción de principio entre símbolos «icónicos» (es decir, las imágenes) y símbolos «discursivos» (es decir, palabras o descripciones). Yo me siento inclinado a considerar que se trata de una actitud razonable aunque, evidentemente, es muy difícil decir en qué consiste esa distinción de principio (véase Goodamn, 1968, y la exposición de Bruner en *Studies in Cognitive Growth*). En cualquier caso, no me voy a ocupar aquí de estos problemas. En seguida va a quedar claro que incluso en el caso de que demos por supuesta la idea de parecido, habrá que atenuar considerablemente el sentido en que los pensamientos *podrían* hacer referencia a sus objetos gracias al parecido con ellos. Es decir, aun cuando la diferencia entre símbolos icónicos y discursivos sea una distinción de principio, la distinción entre las formas en que hacen referencia los símbolos icónicos y discursivos no lo es. Como veremos, no se puede saber *nunca* a qué se refiere un símbolo por el mero hecho de percibirlo, y eso es verdad tanto si el símbolo es icónico como si no lo es.

puesta. Lucienne espera ver en su mano el pliegue del paño, al tener cerrado el puño «como si» la tela estuviera en él. En los momentos más avanzados de la infancia esta primera técnica de representación no desaparece del todo, y es, muy probablemente, el origen de la confusión entre pensar algo y hacerlo (1966, p. 12).

Podríamos preguntarnos, razonablemente, qué clase de argumento podría llegar a esas conclusiones partiendo de premisas como las mencionadas. Indudablemente, muchas de las observaciones de Piaget sugieren la existencia de un período durante el cual el niño se interesa especialmente por los objetos considerados como cosas a manipular; es decir, que los niños muy pequeños suelen atender a las propiedades de los objetos que determinan qué es lo que se puede hacer con ellos. Y existen datos más bien firmes que permiten pensar que, más adelante, los niños se interesan especialmente por las propiedades de los objetos que se pueden reproducir en imágenes —por las propiedades visuales de los objetos, cualquiera que sea el significado preciso de las mismas. Por ejemplo, los niños suelen clasificar las cosas atendiendo a la forma, el color y la simple proximidad, aun cuando sea una forma de hacerlo que a los adultos puede parecer poco natural (véase Vygotsky, 1965); el vocabulario de los niños pequeños suele presentar una preponderancia de palabras referidas a objetos concretos si se las compara con las palabras de abstracciones y relaciones (Brown, 1970), etc. Estas consideraciones pueden hablar en favor de una prominencia especial de los perceptibles en la economía psicológica del niño. Si las cosas son así, nos dicen algo interesante sobre qué es aquello *en* que piensan los niños. Pero de ahí no se deduce que nos digan también algo sobre qué es aquello *con* que piensan los niños. Partiendo de los hechos del tipo de los que acabo de mencionar, Bruner concluye: «...hemos visto que la representación se puede efectuar *en los medios* de símbolos, imágenes y acciones y que cada *forma de representación* puede especializarse en ayudar a la manipulación simbólica, la organización de imágenes o la ejecución de actos motores» (1966, p. 11; el subrayado es mío). La conclusión es, en mi opinión, totalmente gratuita. No se puede, en general, hacer una deducción a partir de *lo que* se representa para llegar a la naturaleza del *vehículo* de la representación. La información sobre las propiedades perceptuales del entorno *podrían,* después de todo, almacenarse como descripciones (es decir, «simbólicamente» en el sentido del término utilizado por Bruner). Por esta razón, demostrar un cambio ontogenético en los rasgos del entorno a que atiende el niño no es más que el primer paso para demostrar la tesis tan radical de que el medio de la representación interna cambia con el desarrollo. Sin embargo, por lo que yo puedo saber, no se ha dado ninguna otra clase de argumentación[11].

Si me he mostrado duro con la base empírica para la existencia de cambios por etapas en las modalidades de representación interna, es porque considero que sería espantoso si los datos nos obligaran realmente a suscribir dicho punto de vista. De

---

[11] Es posible que Bruner piense que los niños utilizan imágenes porque considera que es evidente que no cuentan con ningún medio de representación discursiva; en definitiva, los niños de poca edad no saben hablar. Sin embargo, si es éste el argumento que Bruner tiene «in mente», hay que decir que no es muy convincente. Con la misma razón podríamos decir que los niños pequeños no tienen *imágenes* basándonos en el hecho de que no saben *dibujar*.

hecho, me siento fuertemente inclinado a poner en duda la misma *inteligibilidad* de la sugerencia de que hay una etapa en la que los procesos cognitivos se realizan en un medio que es fundamentalmente no discursivo. No estoy negando, por supuesto, la posibilidad empírica de que los niños puedan utilizar imágenes más que los adultos, o que sus conceptos puedan ser, en un sentido interesante, más concretos que los conceptos adultos. Lo que sí niego, sin embargo, es que la diferencia pudiera ser cualitativa en la forma en que Bruner parece requerirlo. Es decir, creo que no puede haber una etapa en la que las imagénes sean el vehículo del pensamiento en el sentido fuerte de que el pensar sea *identificable* con el imaginar en esa etapa; no, al menos, si las imágenes son representaciones que hacen referencia por el hecho de parecerse. Habría que aclarar todo esto convenientemente.

Imaginemos, *per impossibile,* que los adultos piensan en inglés; es decir, que las oraciones del inglés constituyen el medio en que se llevan a cabo los procesos cognitivos adultos. En ese supuesto, ¿cómo habría que distinguir entre niños y adultos si se mantienen las doctrinas ontogenéticas de Bruner? Es decir, si consideramos que el pensar en inglés es un caso claro de pensar en símbolos, ¿qué es lo que debe figurar como el caso claro correspondiente a pensar en iconos? Una posibilidad es que los niños utilicen un sistema representacional igual que el que utilizan los adultos con la excepción de que los niños tienen *imágenes* donde los adultos tienen *palabras.* Se trata, indudablemente, de una sugerencia coherente; es posible, por ejemplo, imaginarse la concepción de una ortografía jeroglífica para el inglés. De esta manera las oraciones inglesas serían secuencias de imágenes (y no secuencias de fonos) pero todo lo demás seguiría igual. De esta manera hemos atribuido *un* sentido a la propuesta de que el pensamiento de los niños es icónico y el de los adultos es simbólico.

Pero, naturalmente, no es éste el sentido que Bruner tiene presente. Los iconos, en el sentido de Bruner, no son simplemente *imágenes;* son imágenes que se parecen a la cosa a la que hacen referencia. Es decir, no se trata simplemente de que los *símbolos* tengan una *apariencia* diferente a la de los iconos; se trata también de que están diferentemente relacionados con lo que simbolizan. La referencia de los iconos está mediada por el parecido. La referencia de los símbolos está mediada por convenciones. O algo por el estilo [12].

Por eso, el inglés jeroglífico no resultaría demasiado bien. Pero podríamos arreglar las cosas. Es posible imaginar un lenguaje igual que el inglés pero con la única diferencia de que a) las palabras estén sustituidas por imágenes, y b) las únicas imágenes permitidas sean de tal naturaleza que se parezcan a aquello a lo que se refieren las palabras correspondientes. Naturalmente, la capacidad representacional de tal lenguaje sería muy limitada dado que sólo lo podríamos utilizar para referirnos a lo que podemos reproducir en una imagen. Sin embargo, sigue siendo coherente la su-

---

[12] Bruner insiste en el *carácter convencional* de los sistemas representacionales no icónicos (como el inglés), pero, indudablemente, no es su carácter convencional lo que hace que sean no icónicos; el inglés sería un sistema representacional discursivo (es decir, simbólico; es decir, no icónico) aun en el caso de que fuera innato (es decir, no convencional). De hecho, uno de los grandes problemas de la filosofía del lenguaje es dar con una explicación convincente de la relación que existe entre los símbolos y lo que simbolizan. Lo que viene a decir la teoría de Bruner es que los iconos hacen referencia pareciéndose y los símbolos hacen referencia de alguna otra manera —todavía no aclarada—. Esta última afirmación es verdadera, sin duda ninguna.

gerencia de que podría existir este lenguaje, y es una hipótesis coherente proponer que ése es el lenguaje en que piensan los niños. Lo que nos interesa en este ejercicio es que una forma de entender la idea de que los niños piensan en iconos es ésta: Los niños piensan en un lenguaje en el que las *imágenes* (no simplemente jeroglíficos) ocupan el lugar que en los lenguajes naturales ocupan las palabras.

Sin embargo, estoy seguro de que no es ésta la clase de explicación de los procesos mentales de los niños que Bruner trata de recomendarnos. En primer lugar, si la diferencia entre los niños y nosotros fuera sencillamente que nosotros pensamos en algo parecido al inglés standard mientras que ellos piensan en (llamémoslo así) inglés icónico, en ese caso la diferencia entre nosotros y los niños no significaría mucho. Pues aunque el inglés icónico pueda referirse a menos cosas que el inglés standard, ambos pueden expresar algunas de las mismas relaciones semánticas entre las cosas a que se refieren. Después de todo, algunas de estas relaciones se realizan mediante los rasgos gramaticales del inglés standard, y el inglés standard y el inglés icónico tienen la misma gramática. Como en el inglés icónico se podrían expresar, probablemente, el agente, la predicación, posesión, etc., da la impresión de que gran parte de la incapacidad cognitiva que implicaría su utilización sería una pobreza relativa de *vocabulario.* Bruner deja bien claro, sin embargo (1966, cap. 2), que la posibilidad de contar con una estructura gramatical en las representaciones es un rasgo propio de los sistemas representacionales simbólicos (es decir, no icónicos).

Las observaciones precedentes no tienen como intención limitarse a hacer una alabanza de la sintaxis. Lo importante es que podemos dar un sentido al inglés icónico en cuanto sistema representacional precisamente *porque* el paso al inglés icónico deja inalterada la gramática del inglés standard. Una forma de expresarlo sería ésta: en el inglés icónico, las *palabras* se parecen a aquello a lo que se refieren, *pero las oraciones no se parecen a lo que hace que sean verdaderas.* Así, supongamos que, en el inglés icónico, la palabra «John» queda sustituida por una imagen de John y la palabra «green» [=verde] está sustituida por una mancha de color verde. Entonces la oración «John is green» se convertiría (por así decirlo) en una imagen de John seguida de una imagen verde. Pero *eso* no tiene el mismo aspecto que el ser verde; en realidad no se parece a nada. El inglés icónico proporciona una manera de entender la noción de un sistema representacional en que (lo que corresponde a) las *palabras* son iconos, pero no hay manera de entender la noción de un sistema representacional en que lo sean (lo que corresponde a) las *oraciones.* Tampoco yo creo que esto se pueda solucionar de forma útil; la idea de que las oraciones podrían ser iconos no *tiene* interpretación posible. Pero si las oraciones no pueden ser iconos, tampoco podrían serlo los pensamientos.

La estructura del argumento es la siguiente: si el papel que desempeñan las imágenes en un sistema representacional es análogo al papel que desempeñan las palabras en un lenguaje natural, entonces tener un pensamiento *no puede* ser simplemente cuestión de tener una imagen, y esto es cierto tanto si la imagen es motórica como si es icónica y con total independencia de cualquier hipótesis empírica sobre la naturaleza del desarrollo cognitivo. Los pensamientos son las clases de cosas que pueden ser verdaderas o falsas. Son, por tanto, las clases de cosas que se expresan mediante *oraciones,* no palabras. Y, aunque (dejando de lado ciertas consideraciones que mencionaremos más adelante) parece tener cierto sentido imaginar un sistema representacio-

nal en que los equivalentes de las palabras se parezcan a aquello a lo que se refieren, no tiene ningún sentido imaginar un sistema representacional en que ocurra lo mismo con los equivalentes de las oraciones.

Hemos hablado de un hipotético sistema representacional —el inglés icónico— que se distingue del inglés standard en el sentido de que todas las palabras son imágenes pero en el que todo lo demás es idéntico. Hemos señalado que en *ese* sistema representacional existe una relación *no*-icónica entre las oraciones y lo que hace que sean verdaderas. ¿Podemos ofrecer algo mejor? ¿Qué ocurriría si tuviéramos un sistema representacional en que las oraciones fueran iconos de sus condiciones de verdad?

Por ejemplo, ¿qué ocurriría si hubiera un sistema representacional en que la oración «John is fat» [John es gordo] estuviera sustituida por una imagen? Supongamos que la imagen que corresponde a «John is fat» es una imagen de John con una barriga prominente. Pero entonces, ¿cuál sería la imagen que correspondería a «John is tall» [John es alto]? ¿La misma imagen? En ese caso, el sistema representacional no distingue el pensamiento de que John es alto del pensamiento de que John está gordo. ¿Una imagen diferente? Pero John deberá tener una forma u otra en cualquiera de las imágenes que elijamos; por tanto, ¿qué es lo que nos va a decir que tener la imagen es tener un pensamiento sobre la altura de John y no un pensamiento sobre su forma? De la misma manera, una imagen de John es una imagen de John sentado o de pie, o tumbado, o algo indeterminado entre las tres posibles posturas. En ese caso, ¿qué es lo que nos va a decir si tener la imagen es tener el pensamiento de que John es alto, o tener el pensamiento de que John está sentado, o tener el pensamiento de que está de pie, o tener el pensamiento de que está tumbado, o tener el pensamiento de que no se sabe si está sentado, de pie o tumbado? [13].

Existen muchas maneras de abordar esta cuestión. Supongamos que John es gordo y supongamos que el nombre de John es una imagen de John. Pensar en John es tener una imagen que nos presenta a un John gordo. Y pensar que John es gordo es *también* tener una imagen que nos muestre a un John gordo. Pero entonces, ¿cuál es, según esa explicación, la diferencia entre pensar (sólo) en John, por una parte, y pensar que John es gordo, por la otra? [14].

Reconsideremos el camino recorrido. La idea de que los pensamientos son imágenes —o de que eran imágenes cuando éramos muy pequeños— es realmente de una ambigüedad desmesurada. Por una parte, tal propuesta podría equivaler a que debemos identificar tener una imagen con pensar *en* algo, y, por la otra, podría que equivaler a que debemos identificar el tener una imagen con pensar *que* algo. Estas dos proposiciones no corresponden, en absoluto, a lo mismo. La primera equivale a la sugerencia de que las imágenes podrían ser vehículos de *referencia,* mientras que la segunda equivale a la sugerencia de que las imágenes podrían ser vehículos de la *verdad*.

---

[13] Esta forma de argumentación fue propuesta por Wittgenstein (1953). En mi opinión es totalmente convincente.

[14] La solución que parece más clara no es tal solución. Supongamos que pensar en el gordo John *no* implica tener una imagen que represente a John gordo. Sin embargo, la imagen que se tenga deberá representar a John *de alguna manera;* es decir, con unas u otras propiedades. En ese caso, ¿cuál puede ser la diferencia entre pensar simplemente en John y pensar que John tiene esas propiedades?

Así, por ejemplo, si el inglés icónico fuera el lenguaje del pensamiento, el *pensar* en John podría consistir en tener la imagen de John, de la misma manera que, en el uso habitual del inglés común, el *mencionar* a John (hacer referencia a él) podría consistir simplemente en pronunciar su nombre. En este sentido, no es más problemático que haya un lenguaje en que la referencia sea definida en relación con las imágenes que el que haya un lenguaje en que la referencia se defina en relación con las palabras. Supongo que es fácil comprobar en la realidad de los hechos que todos los lenguajes naturales existentes son de la última clase. Pero no veo la forma de elaborar la noción de que podría haber un lenguaje en que la *verdad* se defina en relación con los iconos en vez de con los símbolos; es decir, un lenguaje en el cual las «fórmulas» del sistema sean verdaderas con respecto a aquello a lo que se parecen. El problema estriba *precisamente* en que los iconos son insuficientemente abstractos para ser vehículos de verdad.

En una primera aproximación, lo que puede conseguir un valor de verdad es una atribución de cierta propiedad a algún objeto. Un sistema representacional debe contar, por tanto, con vehículos apropiados para expresar dichas asignaciones. En ese caso, ¿en qué condiciones una representación es adecuada para expresar la atribución de una propiedad a un objeto? Una condición que deberá cumplirse ineludiblemente es que la representación especifique *cuál* de las propiedades se atribuye y cuál es el objeto al que se atribuye. El problema de tratar de valorar la verdad de los iconos es que no nos ofrecen ninguna posibilidad de hacer lo primero. Toda imagen de una cosa, sin excepción posible, representará tal cosa en cuanto que tiene un número indefinido de propiedades; por eso las imagenes se corresponden (o dejan de corresponder) en un número indefinido de formas con las cosas a que se parecen. ¿Cuál de estas correspondencias es la que hace que la imagen sea verdadera?

Pero si las imágenes corresponden al mismo mundo de demasiadas formas diferentes, corresponden también de la misma manera a demasiados mundos diferentes. Una imagen de John con un estómago pronunciado corresponde a que John es gordo. Pero corresponde igualmente a que John está embarazado pues, si John *tiene* ese aspecto cuando está gordo, también lo *tendría,* supongo yo, en caso de que estuviera embarazado. Por tanto, si el hecho de que John esté gordo es razón suficiente para decir que es verdadera una imagen de John con un gran estómago, el hecho de que John no esté embarazado constituye una buena razón para decir que la imagen de John con un estómago prominente es falsa. (Una imagen que corresponde a un hombre que sube una cuesta mirando hacia adelante corresponde igualmente, y de la misma manera, a un hombre que baja la cuesta hacia atrás; Wittgenstein, 1953, p. 139). Por cada razón existente para decir que una imagen es verdadera, tendremos otra razón correspondiente para decir que es falsa. Es decir, no hay ninguna razón para decir ni una cosa ni otra. Las imágenes no son el tipo de cosas que puedan tener valores de verdad.

Téngase presente que los símbolos (a diferencia de los iconos) están libres de estos problemas; este es uno de los sentidos en que los símbolos *son* realmente abstractos. Una imagen de John-gordo es también una imagen de John-alto. Pero la oración «John es gordo» hace abstracción de todas las propiedades de John menos de una: es verdadera si está gordo y sólo si lo está. De la misma manera, la imagen de un hombre gordo corresponde de la misma forma a (es decir, gracias al parecido) a un mun-

do en que los hombres son gordos y a un mundo en que los hombres están embarazados. Pero «John es gordo» hace abstracción del hecho de que los hombres gordos *tienen* una apariencia semejante a la que *tendrían* los hombres embarazados; es verdad en un mundo en el que John está gordo y falsa en cualquier otro mundo.

Vistas en su conjunto estas clases de consideraciones, permiten pensar que no es posible encontrar mucho sentido en la idea de que podría haber un sistema representacional interno en que los iconos fueran vehículos de verdad; es decir, en que tener un imagen sea idéntico a pensar *que* ocurre tal y tal cosa. Pero hemos visto que *podemos* ver cierto sentido en la sugerencia de que hay un sistema representacional interno en que los iconos son vehículos de referencia; es decir, en que pensar *en* tal y tal cosa sea idéntico a tener una imagen. Lo que queremos hacer notar ahora es que hay que protegerse incluso frente a esta concesión.

En inglés icónico el nombre de John es una imagen de John. Por lo cual, si el lenguaje del pensamiento fuera inglés icónico, el pensar en John podría consistir en tener una imagen de John, en el mismo sentido en que, en el inglés normal, el referirse a John podría ser idéntico a proferir «John». Pero ¿qué sentido tiene esto?

Evidentemente, no todas las veces en que se dice «John» se hace referencia a John. Por ejemplo, en el momento en que dejé de escribir a máquina dije «John». Pero no estaba haciendo referencia a nadie; a fortiori, no hacía referencia a John. Podríamos formularlo así: en el caso de los lenguajes naturales, las elocuciones de las expresiones (potencialmente) referenciales sólo consiguen hacer referencia cuando se producen con la intención adecuada. Me es imposible, por así decirlo, hacer una referencia por error; ninguna emisión de voz en que diga «John» puede considerarse como referencia a John a no ser que esté al menos producida con la intención de *hacer* una referencia.

En los lenguajes naturales, dicho en pocas palabras, los vehículos de referencia son elocuciones que se consideran dentro de (es decir que están orientadas a encajar dentro de) ciertos tipos. En los casos paradigmáticos de la referencia a John, digo «John» con la intención de producir una forma verbal, y además de producir una forma verbal utilizada comúnmente para hacer referencia a John. Pero en otras ocasiones en que emito el sonido «John» no se da ninguna de estas circunstancias, y en tales casos (aunque no sólo en esos casos) el hecho de que diga «John» no debe considerarse como referencia a John.

De esta manera, el pronunciar la palabra «John» constituye a veces un caso de hacer referencia a John, pero únicamente cuando el hablante trata de que su conducta se acomode a unas determinadas descripciones; únicamente cuando su elocución está emitida con una determinada intención. Creo que podríamos decir lo mismo, *mutatis mutandis,* del uso de las imágenes en cuanto vehículos de referencia en sistemas como el inglés icónico: si el inglés icónico fuera el lenguaje del pensamiento, podría haber casos en que el tener una imagen de una cosa constituyera un acto de pensar en ella; pero sólo cuando se considera que la imagen entra dentro de determinadas descripciones; sólo cuando se tiene en la forma adecuada. El inglés icónico es, por hipótesis, un lenguaje en que las expresiones referentes son imágenes. Pero incluso en el inglés icónico el parecido no sería condición suficiente de referencia ya que, incluso en el inglés icónico, lo que hace referencia no son las imágenes sino las imágenes ba-

jo determinadas descripciones. En definitiva, el inglés icónico no consigue ser *muy* no-discursivo.

La Figura 4-2 corresponde a una especie de girándula o rueda de fuegos artificiales. Cierre los ojos y hágase una imagen de ella. Si el pensar es formar una imagen de una cosa, y si las imágenes se refieren a todo aquello a lo que se parecen, hemos tenido que estar pensando en un cubo visto desde uno de sus ángulos. La imagen que acabamos de tener se parece, en realidad, a un cubo visto desde esa perspectiva, lo mismo que (y exactamente de la misma manera que) la Figura 4-3 se parece a un cubo visto desde uno de sus lados. Pero, indudablemente, muchos lectores se habrán formado la imagen y *no* habrán pensando en el cubo. El tener la imagen habría constituido un pensamiento en un cubo únicamente en el caso de aquellos lectores que se hubieran formado la imagen y la hubieran considerado de una determinada manera: es decir, han supuesto que el punto del centro era un ángulo del cubo, que las líneas procedentes de dicho punto eran los lados del cubo, etc.

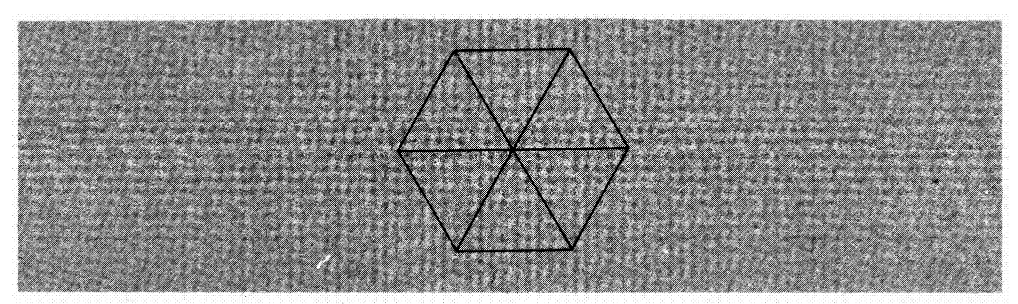
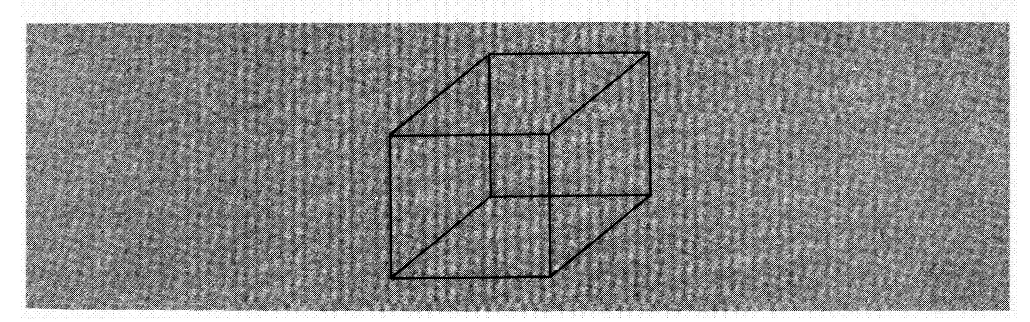


Figura 4-2. Objeto parecido a una girándula. Véase el texto.

La conclusión sería la siguiente: sí, es posible ver cierto sentido en que los niños tengan iconos allí donde nosotros tenemos símbolos; es decir, tengan imágenes donde nosotros tenemos palabras (N.B.: palabras, no oraciones)[15]. Pero no, no es posible



Fiagura 4-3. Cubo esquemático.

[15] Me interesa insistir en que no estoy tratando de *corroborar* la opinión según la cual el pensamiento de los niños es icónico en *cualquier* sentido. Sólo trato de hacer ver lo que representaría una versión coherente de esa forma de pensar. Como habrá quedado ya claro por lo dicho hasta ahora, tal propuesta me parece mucho menos clara que a algunos de los psicólogos que la han suscrito.

ver mucho sentido en la idea de que la relación entre los pensamientos y sus objetos es básicamente diferente en el caso de los niños y en el nuestro. Para que esto tuviera sentido, tendríamos que suponer que las imágenes hacen referencia gracias al parecido mientras que los símbolos lo hacen mediante una convención. (O, como hemos indicado más arriba, algo por el estilo.) Y es evidente que no es eso lo que ocurre. (Las imágenes generalmente no hacen ningún tipo de *referencia*. Pero cuando lo hacen —como, por ejemplo, en el inglés icónico—, lo hacen fundamentalmente de la misma manera que las palabras y frases: es decir, debido a que se acomodan, y a que se considera que se acomodan, a ciertas descripciones.)

Naturalmente, esto no significa que se niegue que las imágenes se parezcan a las cosas de que son imágenes. Lo que se niega es que *parecerse a una cosa* pueda ser condición suficiente para *hacer referencia* a tal cosa, incluso en un lenguaje como el inglés icónico donde las imágenes son las expresiones referentes. De hecho, existe una forma muy adecuada de utilizar una imagen para hacer una referencia: es decir, incrustándola en una descripción. Se puede decir: «Estoy buscando a un hombre que tiene este aspecto...» y señalar una imagen de dicho hombre. Es cierto que, en tal caso, las formas verbales no conseguirían generalmente comunicar una referencia a no ser que la imagen del hombre se parezca al hombre que se busca. Pero, de la misma manera, la imagen no sirve de nada sin la descripción que indique cómo debe ser considerada. Compárense las formas en que se utilizaría la imagen en

$$\text{«Estoy buscando a un hombre que} \left\{ \begin{array}{l} \text{se parece a} \\ \text{viste como} \\ \text{es más alto que} \end{array} \right\} \text{éste... (imagen de un hombre}$$

bajito que lleva una toga)». Lo que contiene la referencia en este caso es la imagen *junto con los «símbolos» que la interpretan.*

En resumen, no veo la forma de interpretar la propuesta de que podría haber un sistema representacional en que el parecido fuera condición *suficiente* para hacer referencia; todavía menos que pueda haber un sistema representacional en que el parecido y la referencia vinieran a ser lo mismo. Dicho brevemente, aunque Bruner estuviera en lo cierto y los *vehículos* de referencia fueran distintos en el caso de los adultos y en el de los niños, los *mecanismos* de referencia —cualesquiera que sean— deben ser muy semejantes en ambos casos.

He estado tratando de echar por tierra dos concepciones sobre las imágenes que han desempeñado un papel influyente y dudoso en la psicología cognitiva: que pensar podría *consistir* en tener imágenes, y que los medios por los cuales las imágenes se refieren a aquello de que son imágenes podrían ser fundamentalmente diferentes de los medios por los cuales los símbolos hacen referencia a lo que denotan. Pero, naturalmente, nada de lo que he dicho equivale a negar que existan imágenes o que éstas desempeñen un papel importante en muchos procesos cognitivos. En realidad, los testimonios empíricos existentes suelen confirmar ambas afirmaciones. Esto es interesante desde el punto de vista de las preocupaciones fundamentales del presente libro. El hecho de que los datos vayan por esa línea constituye una orientación sobre la naturaleza de los recursos representacionales con que contamos. Y el hecho de que contemos con datos, sean como sean, confirma la opinión de que la naturaleza de estos recursos es una verdadera cuestión empírica.

Existen estudios en profundidad de las obras relacionadas con el tema publicadas recientemente (véase, en especial, Paivio, 1971; Richardson, 1969). Aquí nos limitaremos a esbozar uno o dos de los resultados que parecen constituir un fuerte argumento en favor de la realidad psicológica de las imágenes.

1. Si hay imágenes, y si, tal como indica la introspección, tener imágenes se parece mucho a la percepción visual, sería razonable esperar que las tareas experimentales que eliciten imágenes produzcan una interferencia, específica de la modalidad, con otros procesos cognitivos en que interviene la visión. Las tareas que exigen imágenes visuales, por ejemplo, deberían inducir descensos en el rendimiento de las tareas simultáneas que exijan respuestas guiadas visualmente. Una brillante serie de experimentos realizados por Brooks (1968) permite pensar que así ocurre en la realidad. En un caso, se indica a los *S*s que formen una imagen memorística de una figura como la de la Figura 4-4. Luego se les indica que vayan recorriendo dicha imagen siguiendo las flechas e indicando, en cada uno de los ángulos, si se da en uno de los bordes superiores de la figura. (En el caso de la Figura 4-4 las respuestas adecuadas serían: «no, sí, sí, no, no, sí, sí, no, no, no»). Según cuál sea el grupo experimental en que se incluye el sujeto, las respuestas se indican o señalando respuestas afirmativas y negativas ya escritas o mediante alguna forma de gesto no orientado visualmente (como dar golpecitos o decir «sí» o «no»). El resultado pertinente es que el rendimiento es significativamente mejor cuando se trata de sujetos incluidos en los últimos grupos (guiados de forma no visual). Las imágenes visuales se interfieren con las tareas orientadas visualmente.

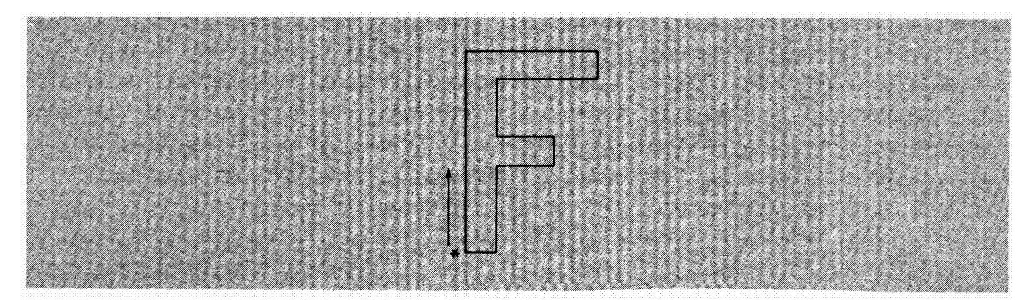


Figura 4-4. Diagrama de estímulo del tipo utilizado por Brooks (1968).

Además, se interfieren de forma selectiva. Brooks mencionaba otro caso en que la tarea de *S* consistía en producir secuencias de síes y noes según cuál fuera la clase de forma de las palabras de una oración memorizada previamente. A un sujeto se le daba, por ejemplo, la oración «Ha llegado el momento de que todos los hombres de bien acudan en ayuda del partido» y se le decía que indicara «sí» por cada palabra que fuera un nombre o verbo y «no» por cada palabra que no fuera ni una cosa ni otra. En esta situación, el efecto de la modalidad de respuesta sobre el rendimiento invertía la relación comprobada en el caso de la imagen visual: el rendimiento era mayor en los sujetos que señalaban o daban golpecitos, mientras que los que hacían las respuestas en forma verbal conseguían resultados más bajos. Parece que las respuestas guiadas visualmente no interfieren demasiado con las imágenes auditivas.

2. Si hay imágenes, y si, como nos indica la introspección, las imágenes se parecen mucho a las reproducciones visuales, debería haber semejanzas demostrables entre los procesos de comparar un objeto con una imagen de dicho objeto y comparar dos objetos que tengan un aspecto semejante. De hecho, se han publicado una serie de experimentos que indican que así ocurre en la realidad. (Véase, por ejemplo, Cooper y Shephard, 1973). El estudio paradigmático es el de Posner, Boies, Eichelman y Taylor (1969).

En primer lugar, es posible demostrar que existe una diferencia firme en la velocidad con que los sujetos pueden emitir juicios sobre la identidad de tipos en el caso en que los ejemplos son físicamente *semejantes,* por una parte, y en el caso en que son físicamente *diferentes,* por la otra. Así, por ejemplo, se presentan a los *S*s muestras taquistoscópicas formadas por dos letras y se les dice que respondan «sí» si son las mismas letras y «no» si son diferentes. En esta situación, los *S*s son más rápidos cuando los miembros de los pares positivos (es decir, los pares donde la respuesta correcta es «sí») son de la *misma caja* (por ejemplo *PP* o *pp*) que cuando son de *diferente caja* (por ejemplo *Pp* o *pP*).

Supongamos ahora que se cambia el paradigma. En vez de presentar al *S* dos letras dentro de la modalidad visual, le presentamos en primer lugar una designación auditiva, y luego *una sola letra visual* para emparejarla con la descripción auditiva. Así el sujeto oiría «*P* mayúscula» y luego vería *P* (en cuyo caso la respuesta debería ser «sí») o *p* o *Q* o *q* (en todos estos casos la respuesta correcta sería «no»). Resulta que el rendimiento de *S* en esta situación depende críticamente de la longitud del intervalo entre el estímulo auditivo y el visual. Los sujetos que reciben el estímulo visual inmediatamente después del estímulo auditivo manifiestan latencias de respuesta comparables a las de los pares de letras presentados visualmente cuyos miembros *difieren* en cuanto al tipo de caja (mayúscula/minúscula). Sin embargo, si se aumenta el intervalo entre uno y otro estímulo hasta aproximadamente 0,7 segundos, las latencias de respuesta *disminuyen* y se aproximan a las de los pares de letras presentados visualmente cuyos miembros son *idénticos* en cuanto al tipo de caja. No es ineludible, pero sí enormemente natural, suponer que lo que ocurre durante los 0,7 segundos del intervalo entre uno y otro estímulo es que el sujeto construye una imagen de la letra que se adapte a la descripción auditiva, y que es esta imagen la que se empareja con la reproducción visual. Si esto es así, y si, como hemos supuesto, el emparejar las imágenes y las cosas es fundamentalmente semejante a emparejar cosas que tienen una apariencia semejante, tenemos una especie de explicación de la convergencia conductual entre los *S*s que juzgan la relación entre pares de letras que *ven,* y los *S*s que juzgan la relación existente entre pares de letras en los que una de ellas sólo tiene una descripción auditiva.

Los estudios que acabamos de mencionar no son en absoluto las únicas posibilidades para la investigación empírica de la realidad psicológica de las imágenes mentales[16]. Consideremos otra posible línea argumental.

---

[16] Quizá el descubrimiento más impresionante sea que la percepción estereóptica de la profundidad se puede producir imponiendo una imagen eidética a un estímulo visual. (Véanse los resultados sorprendentes citados por Stromeyer y Psotka, 1970. El lector interesado por una exposición más general del eidetismo puede consultar Haber, 1969). Parece difícil negar que el hecho de tener imágenes sea equivalente al de

Los símbolos discursivos, como indicaba Bruner, se despliegan en el tiempo. O, más bien, eso es lo que ocurre con *algunos* símbolos discursivos (a saber, las oraciones habladas). Las imágenes (y las oraciones escritas) se despliegan en el espacio. Puede que haya convenciones para determinar el orden en que se recupera la información a partir de una imagen (como en ciertas clases de dibujos didácticos que «cuentan una historia» y están pensados para ser examinados en un cierto orden) pero, en general, no hace falta que las haya. En principio, toda la información está accesible de forma simultánea y se puede leer en el orden que elija el observador[17].

Supongamos que los sujetos *pueden* emplear imágenes mentales para reproducir la información pertinente a la realización de una tarea experimental, y supongamos que las imágenes mentales son semejantes a las imágenes reales. En ese caso, sería posible prever que los *S*s que pueden utilizar imágenes deben disponer de una libertad considerable en cuanto al orden en que comunican la información que presentan sus imágenes, mientras que los *S*s que utilizan formas discursivas de representación (por ejemplo, oraciones) deberían estar relativamente limitados en cuanto al orden en que se puede tener acceso a la información. Por poner un caso extremo, imaginemos un experimento en que el sujeto se encuentra con un triángulo rojo y luego se le pregunta por lo que acaba de ver. Los *S*s que han almacenado una *imagen* deberían responder prácticamente con la misma rapidez a las preguntas «¿Era rojo?» y «¿Era triangular?» Los *S*s que almacenaron la oración «Era un triángulo rojo» deberían responder con mayor rapidez a la primera pregunta que a la segunda* [18].

Pero, por desgracia, tal como son las cosas, hay que decir que se trata de un experimento en gran parte de tipo especulativo *(gedanken);* lo menciono fundamentalmente como una ilustración más de las técnicas que se podrían utilizar para someter a comprobación experimental las hipótesis sobre la naturaleza de las representaciones internas. Sin embargo, conviene señalar que es precisamente esta interpretación la propuesta por Paivio (1971) para explicar las diferencias en cuanto al orden-de-informe presentadas por los sujetos de un experimento de Haber (1966). Paivio indi-

---

percibir cuando es posible producir ilusiones perceptuales características cuyos objetos son imágenes y no perceptos. Conviene recordar, en ese sentido, que hace ya tiempo que sabemos que hay circunstancias en que los sujetos se pueden ver inducidos a confundir las imágenes (*no*-eidéticas) con los perceptos (Perky, 1910; Segal y Gordon, 1968).

[17] Este punto tiene relación con un comentario de Kant en la *Crítica de la razón pura.* Kant distingue entre secuencias temporales «subjetivas» y «objetivas», siendo las segundas, pero no las primeras, independientes de las estrategias de exploración del sujeto que percibe. Así, podemos elegir examinar la fachada de un edificio desde el pórtico al frontón. Pero como todas las partes del edificio son contemporáneas, podríamos haber optado por proceder en sentido inverso. Los hechos que constituyen una secuencia objetiva, por el contrario, sólo se pueden explorar en una única dirección. Lo mismo podría decirse, *mutatis mutandis,* del contraste que existe entre recuperar la información de las imágenes y de las oraciones habladas.

* En inglés el orden de colocación de adjetivo y sustantivo es inverso al español: *red* [=rojo] *triángulo* [= triángulo].

[18] Los *S*s que almacenaron, por ejemplo, la oración «Era un triángulo y era rojo» deberían reflejar, lógicamente, la asimetría contraria. Lo importante es que habría que asociar *una u otra influencia en el orden del informe* con una forma de representación discursiva, mientras que los que actuaban con imágenes deberían estar relativamente libres de tales influencias. Si resultara que los *S*s que afirman que tienen imágenes fueran los que presentan influencias relativamente débiles en el orden del informe, tendríamos una razón para tomarnos en serio la hipótesis de que *están* utilizando imágenes.

ca que «aunque no se han comprobado independientemente las consecuencias del presente análisis utilizando las tareas perceptuales adecuadas, hay pruebas de orígenes diversos que son consistentes con esta hipótesis» (p. 130).

Lo que venimos diciendo debería hacernos pensar que es posible tratar de la existencia y funcionamiento de las imágenes como auténticas cuestiones experimentales y que en los experimentos se pueden utilizar técnicas más sutiles que una elemental llamada a la introspección. Quizá esto resulte sorprendente al lector de formación filosófica, pues en estos últimos tiempos ha estado en boga la tendencia a considerar que la no-existencia de las imágenes era demostrable a priori. Antes de dar por finalizada esta cuestión, puede resultar interesante hacer una digresión para ver qué es lo que se puede decir en favor de un punto de vista tan poco plausible.

Dennett (1969) ha formulado sucintamente lo que parece constituir el principal problema filosófico de las imágenes.

Pensemos en el tigre y en las listas de color de su piel. Puedo soñar, imaginar o ver un tigre listado, pero ¿debe tener el tigre de mi experiencia un número concreto de listas? Si ver o imaginar es tener una imagen mental, entonces la imagen del tigre —para obedecer a las reglas de las imágenes en general— *debe* revelar un número determinado de listas y debe ser posible precisarlo con preguntas como «¿más de diez?», «¿menos de veinte?». Sin embargo, si ver o imaginar tiene un carácter descriptivo no es necesario que las preguntas tengan una respuesta precisa. A diferencia de la instantánea de un tigre, la descripción del mismo no tiene ninguna necesidad de detenerse en el número de listas; es posible que la descripción se limite a decir «numerosas listas». Naturalmente, en el caso de que se vea realmente un tigre, muchas veces será posible examinarlo detenidamente y contar las líneas de color, pero en ese caso se cuentan listas reales, no las franjas de una imagen mental (pp. 136-137).

Hay filósofos que parecen afirmar que este tipo de argumentación constituye algo semejante a una *demostración* de que no hay imágenes mentales. Si tienen razón, su afirmación es embarazosa ya que, como hemos visto, existen ciertas pruebas empíricas convincentes que indican que lo que ocurre al imaginar se parece mucho a la reproducción de imágenes y muy poco a la descripción. Además, la plausibilidad introspectiva de la teoría de la imagen es enorme, por lo que si el tigre listado demuestra realmente lo que se dice que demuestra nos quedamos sin una explicación de las introspecciones o de los datos experimentales. Cualquier teoría es mejor que no tener ninguna; es evidente que debemos tratar de echar por tierra el argumento del tigre listado, si podemos.

Existen, en mi opinión, al menos tres formas de intentarlo. No estoy diciendo que cualquiera de estos contraargumentos sea definitivo, pero considero que, entre todos, nos indican que los tigres listados no constituyen un argumento irrefutable en contra de las imágenes. Dado el carácter persuasivo de los argumentos a posteriori en favor de las imágenes, debería ser suficiente con eso.

En primer lugar, podríamos tratar simplemente de rechazar lo que da por supuesto el argumento del tigre. Es decir, podríamos argumentar diciendo que *existe* una respuesta precisa a la pregunta «¿Cuántas franjas tiene el tigre-imagen?» pero que, debido a que nuestras imágenes son lábiles, generalmente no podemos mantenerlas el tiempo suficiente como para poder contarlas. En favor de este punto de vista hay que decir a) que, introspectivamente, parece plausible a muchas personas que afir-

man tener imágenes (si alguien no lo cree, puede preguntar a otros)[19]; b) hace que las imágenes mentales cotidianas sean cualitativamente del mismo tipo que las imágenes eidéticas, de las que incluso Dennett admite que «el sujeto *puede* contar y enumerar los detalles» (p. 137); c) este punto de vista resulta menos difícil de aceptar que la afirmación contraria: que lo que ocurre cuando pienso que estoy reproduciendo algo en imágenes es que, de hecho, me lo estoy describiendo[20].

Este es, creo yo, el tipo de afirmación que los filósofos más profundos suelen considerar ingenua; quizá porque se dejan impresionar por este tipo de argumentación: «Se considera que tener imágenes forma parte del proceso de percepción. Ahora bien, si hay que percibir las mismas imágenes (explorarlas, etc.) para recuperar la información que contienen, es indudable que hemos dado el primer paso en un retroceso que acabará obligándonos a postular imágenes sin número y una serie interminable de perceptores que las miren». Sin embargo se trata de un argumento de poco valor. Supone, sin ninguna justificación, que si el recuperar la información del entorno externo exige que se tenga una imagen, el recuperar la información de una imagen debe exigir también que tengamos una imagen. Pero ¿por qué hay que hacer esta suposición? Además (y más en relación con el punto que nos ocupa), aun cuando el argumento tuviera valor no lo tendría aquí. Lo más que podría demostrar es que las imágenes no desempeñan un papel determinado en la percepción (es decir, que la percepción de una cosa no puede requerir siempre y en todo lugar la formación de una imagen de dicha cosa). No demuestra nada sobre si tener y explorar una imagen puede desempeñar un papel en *otros* procesos mentales (tales como, por ejemplo, comparar, recordar o imaginar cosas).

La segunda afirmación que nos interesaría hacer sobre el argumento de los tigres con listas es ésta: hay que decir, sin más, que no es cierto que una imagen de un tigre listado deba ser determinada con respecto a descripciones tales como «tiene *n* listas»[21]. Por supuesto, el *tigre* debe tener un número preciso *n* de listas (prescindiendo de los problemas sobre la individualización de las listas concretas), pero hay casos de todo tipo en los que la imagen de un tigre con *n* listas no representa un número determinado de listas. El principal (pero no el único problema) es su carácter borroso[22].

Lo que *es* cierto, lo que sí se deduce de lo que Dennett llama «las reglas de las imágenes en general» es que si lo que tenemos es una imagen, deberá haber necesa-

---

[19] *No quiero* entrar en la cuestión de si la introspección es infalible; pero parece un tanto perverso afirmar que lo que dice la introspección está siempre, eo ipso, equivocado. Parece lógico considerar que las opiniones del sujeto sobre qué es lo que hace tiene el mismo derecho a ser respetadas que las mías o las del lector o las del experimentador.

[20] Todavía cuesta más creer que lo que ocurre en los casos característicos de la *percepción* de una cosa sea significativamente semejante a lo que ocurre en los casos característicos de descripción de la misma. La afirmación es pertinente, ya que la forma natural de considerar las imágenes consiste en pensar que imaginar una cosa es estar en un estado psicológico cualitativamente similar al estado en que se estaría si es estuviera percibiendo la cosa. Por lo tanto, si imaginar es como describir, también lo debe ser el percibir.

[21] Parto del supuesto de que una imagen es *determinada según una descripción* si y sólo si la afirmación de que la imagen se acomoda a la descripción tiene un determinado valor de verdad.

[22] Pensemos en la fotografía desenfocada de una página impresa. Hay una respuesta precisa a la pregunta «¿Cuántas letras hay en la página?». ¿Es necesario que haya una respuesta precisa a «cuántas imágenes de letras hay en la fotografía?».

riamente *una* descripción visual según la cual sea determinada. En el caso de la foto de un periódico, por ejemplo, la descripción pertinente es la que especifica una «matriz de grises»; la atribución de un valor de blanco o negro a cada uno de los numerosos puntos que comprenden la imagen. Según mis conocimientos, este es el *único* tipo de descripción visual en que las fotos de un periódico son *siempre* determinadas. El que una foto semejante sea también determinada de acuerdo con alguna *otra* descripción visual (como, por ejemplo, tiene *n* franjas) dependerá de cuál es el tema de la foto, el ángulo desde el que se tomó, la calidad de la resolución, etc.

Si lo que decimos es cierto, se puede afirmar que el argumento del tigre listado es mucho más flojo de lo que parecía al principio. Lo que demuestra dicho argumento, *en el mejor de los casos,* es que hay *ciertas* descripciones visuales según las cuales las imágenes mentales *no son* plenamente determinadas. Pero lo que habría que demostrar para probar que las imágenes mentales no se atienen a «las reglas de las imágenes en general», es decir, para probar que no son imágenes, es que no hay *ninguna* descripción visual según la cual *estén* totalmente determinadas. Es claro que de las observaciones hechas por Dennett no se puede deducir una conclusión tan tajante[23].

Lo tercero que queremos decir contra el argumento del tigre es que es más dogmático sobre la distinción entre imágenes y descripciones de lo que hace falta. Una imagen paradigmática (por ejemplo una fotografía) es *no-discursiva* (la información que transmite aparece reflejada más que descrita) y *gráfica* (se parece a su tema). Lo que interesa en este momento, sin embargo, es que existe una gama indefinida de casos intermedios entre fotografías y párrafos. Estos casos intermedios son imágenes según unas determinadas descripciones; transmiten *cierta* información discursivamente y *otra* información gráficamente, y se parecen a sus temas únicamente en relación con aquellas propiedades que aparecen representadas gráficamente. En concreto, son determinadas según las mismas descripciones visuales que sus temas únicamente en relación con estas propiedades[24].

Quizá consigamos aclarar la cuestión por medio de un ejemplo. Dennett dice: «Consideremos la versión cinematográfica de *Guerra y paz* y la obra original de Tolstoy; la versión cinematográfica es enormemente detallista y en cierto sentido es imposible que sea *fiel* al texto de Tolstoy, pues el «cuadro pintado» por el novelista no precisa muchos detalles que la película no puede dejar de precisar (por ejemplo, el color de los ojos de cada uno de los soldados que aparecen en la película» (1969, p. 136). Pero, además, existen imágenes que no son fotografías. Como ejemplo se pueden citar los *mapas*. Los mapas son gráficos en relación con parte de la informa-

---

[23] En mi exposición se presupone la cuestión de qué es lo que debe entenderse por descripción «visual». Sin embargo, ocurre lo mismo en el argumento del tigre listado pues, probablemente, sólo en el caso de las descripciones visuales se deduce de «las reglas de las imágenes en general» que las imágenes deban ser determinadas.

[24] No se trata siquiera de que las imágenes relacionadas con una descripción sean necesariamente gráficas en relación con toda la información respecto a la cual son no-discursivas. Uno de los motivos de confusión al hablar de las imágenes está en considerar que «no-discursivo» y «gráfico» son coextensivos. Así, la línea del globo terráqueo que indica dónde se encuentra el ecuador transmite la información de forma no discursiva. Pero no se parece al ecuador. Estos casos indican lo impreciso que es el contraste no analizado entre imágenes y descripciones. En este contexto, estoy utilizando los materiales con que cuento, pero un trabajo serio de investigación dentro de este campo debería tratar de precisar (y quizá, en último término, abandonar) el marco de las distinciones que he utilizado aquí.

ción que comunican; cuando se orientan adecuadamente, aparecen reflejadas las relaciones geográficas. Pero no son, o pueden no ser, gráficos en relación con otras informaciones. Las densidades de población o la altitud sobre el nivel del mar pueden aparecer representados por colores, y en ese caso tenemos que recurrir a una explicación de signos convencionales para determinar qué es lo que quiere decir la imagen.

Dicho más brevemente, como las imágenes sometidas a descripción son imágenes, suelen ser gráficas en relación directa con un conjunto determinado de propiedades visuales, y, por lo tanto, es evidente que estarán determinadas directamente por el conjunto de propiedades que representan gráficamente. Pero como, en parte, es la descripción lo que determina *de* qué es imagen la mencionada imagen, las propiedades en relación con las cuales la imagen debe ser determinada pueden tener poco en común con las propiedades visuales de lo que la imagen represente de forma arbitraria. Las imágenes con descripciones comparten su carácter no discursivo con las imágenes *tout court.* Lo que tienen en común con las descripciones es que no es necesario que se parezcan demasiado a lo que representan.

Y ya estamos en condiciones de decir qué tiene que ver todo esto con las listas de color del tigre. Supongamos que lo que se visualiza al imaginar un tigre puede ser desde un retrato a escala natural del tigre (en el caso del eidetista) hasta una especie de figura esquemática pasajera (en el caso de los que no tienen mucha imaginación, como yo). Lo que hace que mi figura esquemática sea una imagen de un tigre no es que se le parezca excesivamente (tampoco mis dibujos se parecen demasiado a los tigres) sino, más bien, que es *mi* imagen, por lo cual soy yo quien puede decir de qué es imagen. Mis imágenes (y mis dibujos) están conectadas con mis intenciones, en cierta manera; yo las *considero* como imágenes de un tigre para cualquier tarea que me vea obligado a realizar. Como mi imagen mental *es* una imagen, habrá ciertas descripciones visuales según las cuales sea determinada; por eso, habrá ciertas preguntas cuyas respuestas pueda «leer al vuelo» en la representación visual[25], y cuanto más gráfica sea la representación mayor será el número de esas preguntas. Pero, en el caso de una imagen determinada, puede haber, de forma arbitraria, un número de propiedades visuales que no aparecían reflejadas sino, por así decirlo, transmitidas por la descripción a que trata de adaptarse la imagen. La imagen, ipso facto, no estará determinada en relación con estas propiedades. Volvemos así, por una ruta diferente, a la conclusión debatida anteriormente: para demostrar que las imágenes mentales van contra «las reglas de las imágenes en general», habría que demostrar no sólo que son indeterminadas según una descripción visual u otra, sino más bien que no son determinadas en ningún tipo de descripción visual. *Quizá* sea posible demostrar esto pero lo dudo, y desde luego el argumento del tigre no lo consigue.

Todo esto nos lleva a formular ciertas especulaciones plausibles sobre la forma en que las imágenes se pueden integrar con las modalidades discursivas de la representación interna. Si recordamos el experimento de Posner *et al.,* mencionado más arriba, nos daremos cuenta de que hay dos procesos psicológicos que parecen deducirse de la

[25] Probablemente, las personas se toman la molestia de construir imágenes en las tareas de memoria porque las imágenes permiten «leer al vuelo» algunas informaciones. Una anécdota psicológica clásica es la del hombre que no sabe cuántas ventanas tiene su casa a no ser que se construya una imagen de la casa y luego empiece a contar.

explicación de los resultados que proponen los autores. En la primera fase, se construye una imagen en conformidad con una descripción. En la segunda, la imagen se compara con un estímulo en orden a una identificación perceptual. De esta manera, la explicación implica (en conformidad con lo que nos sugiere el sentido común) que tenemos facultades psicológicas que pueden construir imágenes que manifiestan la información transmitida discursivamente por las descripciones correspondientes; es decir, facultades que nos permiten construir imágenes *a partir de* descripciones. El experimento demuestra que en ciertos tipos de tareas el rendimiento se ve facilitado cuando se tiene la información expresada en forma de imagen. (En efecto, la utilización de la imagen más que de la descripción hace posible que el sujeto realice la tarea de categorización perceptual en forma paralela más que serial; puede comprobar *al mismo tiempo* de qué letra se trata y si es mayúscula o minúscula.)

Estas observaciones sobre el experimento de Posner encajan muy bien con la idea de que las imágenes acompañadas de descripciones son muchas veces vehículos de representación interna. En la medida en que las imágenes mentales se construyen *a partir de descripciones,* éstas pueden servir para determinar de qué son imágenes las imágenes y cómo se deben interpretar sus propiedades. He aquí, pues, un esquema general de la idea que he tratado de desarrollar:

1. Algunas conductas se ven facilitadas cuando la información relevante para la tarea se representa en forma no discursiva (por ejemplo, cuando se representa en forma de imagen).
2. Una de nuestras facultades psicológicas funciona con la misión de construir imágenes que estén en conformidad con las descripciones. Es decir, tenemos acceso a un sistema computacional que tiene como input una descripción y ofrece, como output, la imagen de algo que se ajusta a la descripción. La explotación de este sistema es, probablemente, sensible a nuestra estimación de las características de demanda de la tarea que tenemos entre manos.
3. La imagen que se produce puede ser muy esquemática, pues la forma en que se *considera* la imagen —cuál es el papel que desempeña en el procesamiento cognitivo— está determinada no sólo por sus propiedades figurativas sino también por el carácter de la descripción con que se empareja. Hemos visto que este punto es importante para evaluar el argumento del tigre listado. Podríamos añadir ahora que supone un cierto avance en cuanto forma de responder a uno de los argumentos empíricos utilizados frecuentemente para atacar a quienes se toman muy en serio las imágenes mentales.

Los psicólogos que creen que las imágenes no pueden desempeñar un papel muy importante en la representación interna insisten muchas veces en el carácter idiosincrásico de las imágenes de que hablan los sujetos (véase, por ejemplo, Brown, 1958). Evidentemente, el contenido de las imágenes varía mucho de una persona a otra, y sería perfectamente posible que una imagen determinada diera lugar a distintas representaciones en tareas computacionales diferentes (lo que figura como imagen de un pato en un sentido puede figurar en otro como imagen de un conejo). Lo importante, en este contexto, es que si las imágenes mentales son imágenes que corresponden a descripciones, sus indiosincrasias podrían influir muy poco en el papel que desempeñan en los procesos cognitivos. Supongamos que mi imagen de un triángulo es

isósceles y la del lector es escaleno. No es necesario que esto influya en la forma de utilizar las imágenes para razonar sobre los triángulos en la medida en que estemos de acuerdo en la forma en que se deben interpretar las imágenes; por ejemplo, en la medida en que estemos de acuerdo en que deben representar *cualquier* figura cerrada de tres lados cuyos lados sean líneas rectas.

Esta observación es ya tradicional. Los empiristas ya la expusieron, aunque muchas veces se ha pasado por alto la importancia de sus opiniones. Así, Hume admitía la intuición de Berkeley según la cual las imágenes no se pueden parecer a los referentes de las ideas abstractas, pero afirmaba que hay un sentido en el cual el tener una idea abstracta podría ser idéntico a tener una imagen. Dice Hume: «La imagen mental es sólo imagen de un objeto concreto, aunque la aplicación que hacemos de ella en nuestro razonamiento sea la misma que si fuera universal» (1900 ed., p. 28). Considerado de una manera, esto equivale al abandono de la teoría imaginística del pensamiento, pues los vehículos de la representación interna se consideran (no imágenes *tout court* sino) imágenes según una u otra interpretación; lo que hemos llamado imágenes correspondientes a descripciones. Lo que se ha abandonado, en concreto, es la doctrina de que las imágenes mentales se refieren a aquello a lo que se parecen y se parecen a aquello a lo que se refieren. Pero, si lo consideramos desde el punto de vista contrario, lo que quiere decir Hume es que el abandono de las teorías del parecido es compatible con la conservación de la idea de que (algunas de) las representaciones internas son, o pueden ser, no-discursivas. La importancia de distinguir entre estas dos propuestas —y la incapacidad de muchos filósofos y psicólogos de hoy día para verlo— ha sido uno de los temas principales de nuestra exposición.

Hasta el momento no hemos conseguido más que un esbozo de una teoría: las cuestiones que deja abiertas son más interesantes que aquellas a las que trata de dar respuesta. Por ejemplo, suponiendo que exista lo que llamamos imágenes mentales, ¿hay alguna razón para suponer que desempeñen en la representación interna un papel que no sea simplemente marginal? ¿Qué clases de tareas se ven facilitadas por la disponibilidad de representaciones no-discursivas? ¿Qué tienen las representaciones no-discursivas que hace que sean útiles en dichas tareas? ¿Cuánta libertad tenemos para optar por la representación no-discursiva en determinados casos? ¿Cuáles son los mecanismos mediante los cuales se construyen imágenes a partir de las descripciones?[26]. Por encima de todo, sería interesante saber si *todas* las imágenes mentales se engendran a partir de descripciones o si ciertos procesos psicológicos son, por así decirlo, no-discursivos desde el principio hasta el final[27]. Si, por ejemplo, utilizo imáge-

---

[26] Pueden conseguirse algunas pistas mediante el examen de las rutinas de computación «que convierten lo digital en analógico». Dicho examen habla en favor de la posibilidad de dispositivos psicológicamente reales que reproduzcan las descripciones en imágenes, pues existen ya máquinas que pueden realizar tales funciones. Véase Sutherland (1970).

[27] Por las razones que acabamos de mencionar, supongo que en la medida en que las representaciones internas son imágenes deben ser imágenes-relacionadas-con-una descripción. Lo que considero una cuestión empírica abierta es la de los mecanismos mediante los cuales se relacionan las descripciones y las imágenes. Una forma de relacionarlas —la que hemos esbozado más arriba— consistiría en generar las imágenes *a partir de* las descripciones. Lo que nos preguntamos ahora es si hay otras formas y, en ese caso, cuáles son.

A propósito de esto, quizá sea conveniente señalar que existen semejanzas entre lo que vengo diciendo

nes para recordar la apariencia o el olor de una cosa, ¿recurro invariablemente a una información que estuvo representada discursivamente en algún momento de su historia? Lo que había almacenado Proust, ¿era una *descripción* de cómo saben las magdalenas con té? ¿O existen mecanismos psicológicos mediante los cuales se establecen y despliegan los engramas no discursivos? Ciertamente, las cantidades enormes de información que se manipulan en ciertas tareas en las que se implican imágenes hacen poco probable que la información representada atravesara una etapa de codificación digital. En cualquier caso, la discusión ha vuelto a plantearse en un terreno de la investigación psicológica que es claramente empírico, y ahí es donde yo trato de dejarla. El lector interesado en el tema puede consultar Pribram (1971) y Penfield y Roberts (1959).

Muchos procesos psicológicos son computacionales; implican esencialmente la transformación de la información que el entorno perceptual (o genético) del organismo pone a su disposición. Pero la información debe estar representada de alguna manera, y ciertas formas de representación son mejores que otras; es decir, están mejor adaptadas a la tarea en que está inmerso el organismo. El problema biológico con que nos encontramos al diseñar la psicología de los organismos consiste, por lo tanto, en asegurar, en la medida de lo posible, que las modalidades de representación estén emparejadas óptimamente con las distintas clases de tareas. Los seres humanos son una forma de solución a este problema, y lo mismo ocurre, supongo yo, con cualquier otro organismo que tenga vida mental.

Lo que me he propuesto en este capítulo ha sido hacer ciertas observaciones sobre el tipo de solución que constituyen los humanos. La clave parece estar en la flexibilidad. Los seres humanos parecen tener acceso a distintas modalidades de representación y pueden ejercer un control racional sobre los tipos de representación que emplean. Es decir, la forma en que se explotan los recursos representacionales disponibles en un caso determinado depende de lo que el agente considera que son las exigencias de la tarea que tiene entre manos. El despliegue eficaz de las capacidades computacionales constituye de por sí un problema computacional, y parece que los seres humanos están bastante bien equipados para resolverlo.

Quizá este resultado parezca decepcionante a los psicólogos que tienen prisa por dar con soluciones. Muchas veces se trata de realizar experimentos que pongan en juego los mecanismos involuntarios de la cognición; los reflejos intelectuales, por así decirlo, con que la mente responde, de grado o por fuerza, a la tarea. Pero lo que suele encontrarse son simplemente las estrategias locales, orientadas a un objetivo concreto, que elaboran los sujetos en orden a cumplir eficientemente las instruccio-

---

sobre la forma en que podrían desplegarse las imágenes en las tareas de reconocimiento y las llamadas teorías de «análisis por síntesis» de la categorización perceptual. La idea de estas teorías es, precisamente, que las representaciones —en realidad, plantillas— se engendran a partir de las descripciones y luego se emparejan con el input que necesita ser categorizado. De esta manera, la descripción a partir de la cual se engendra la plantilla proporciona el análisis perceptual del input. Una característica atractiva de estos modelos es que proporcionan una reserva infinita de plantillas, en la medida en que las reglas de formación de las descripciones son iterativas. (Puede verse una exposición más detallada en Halle y Stevens, 1964; Neisser, 1967.) Dudo mucho que el análisis por síntesis pueda producir algo parecido a una teoría general de la percepción, pero es muy plausible que dichos mecanismos intervengan, *inter alia,* en la perpcepción.

nes que se han dado. De esta manera, lo que el experimento revela fundamentalmente es la capacidad del sujeto para adivinar los objetivos del experimentador, y su disposición, en general, a hacer todo lo que está de su parte para conseguirlos. (Las publicaciones más recientes sobre el «condicionamiento verbal» constituyen un ejemplo muy claro de lo que decimos. Véase Brewer, de próxima aparición; Dulaney, 1968.)

Se dice que la ciencia debe explicar las uniformidades que están por debajo de la confusión superficial de los acontecimientos. Por eso, resulta deprimente profundizar por debajo de los complejos y cambiantes recursos cognitivos que aportan los sujetos humanos a la tareas de resolución de problemas, para encontrar únicamente, una y otra vez, nuevos estratos de recursos cambiantes y complejos. Pero la *simple* depresión no sería la respuesta más acertada. Después de todo, no deja de tener interés que nuestras capacidades cognitivas estén estratificadas en la forma en que lo están. Por el contrario, parece cada vez más claro que una teoría de la organización racional de los recursos computacionales será una parte significativa de cualquier explicación de por qué hacemos tan bien lo que hacemos. Y si lo que he expuesto en este capítulo no está descaminado, la organización de los medios de representación constituye una parte de lo que tendrán que considerar tales teorías.

A la larga —*muy* a la larga— tendremos que abordar el tema de los reflejos involuntarios de la cognición. En ese largo plazo, necesitamos una teoría, no sólo de un proceso racional u otro, sino de la racionalidad per se. Como ha señalado Dennett, no contaremos con tal teoría mientras nuestras explicaciones contengan palabras «"mentalistas" como "reconocer" y "calcular" y "creer" [pues estas explicaciones] presuponen el mismo conjunto de capacidades —cualesquiera que sean las capacidades que contribuyen a formar la inteligencia— que deberían explicar» (sin publicar). Sin embargo, parto del supuesto de que todos nosotros somos psicólogos que nos movemos en el plazo medio y que nuestros objetivos provisionales son bastante menos ambiciosos que la eliminación en masa de los predicados intencionales de las explicaciones psicológicas. Dentro de nuestros fines más modestos, no se trata de eliminar la racionalidad, sino sencillamente de demostrar cómo se estructura la racionalidad. Teniendo esto en cuanta, puede ser suficiente que las teorías psicológicas manifiesten las formas en que los procesos racionales dependen unos de otros. Si la mente es, después de todo, un mecanismo, estas teorías no podrán decir la última palabra en psicología. Pero el problema con que nos enfrentamos ahora no es el de decir la última palabra. Nuestro problema es encontrar algo —casi cualquier cosa— que podamos decir que es *verdad.*

Conclusión

ALCANCE Y LIMITES

Siempre pienso que cuando se tiene la sensación de haber llevado una teoría demasiado lejos es cuando ha llegado el momento de llevarla todavía un poco más lejos.
¿Un poco? ¡Pero bueno! ¿Es que te estás haciendo viejo?

MAX BEERBOHM

Como el resto de las ciencias, la psicología no empieza desde cero. Esto quiere decir no solamente que los psicólogos heredan de la cultura en sentido amplio un legado de creencias, de explicaciones y teorías presistemáticas y sólo parcialmente articuladas, sobre la forma en que funciona la mente, sino también que las cuestiones sobre la mente que definen su campo son, en un primer momento, simplemente las que la investigación informal ha llegado a plantear pero no a resolver: ¿Cómo aprendemos? ¿Cómo percibimos? ¿Qué es el pensamiento? ¿Cómo se expresan en palabras los pensamientos?

Las diferentes escuelas psicológicas se distinguen, *inter alia,* por las actitudes que tienen hacia este legado. El conductismo, por ejemplo, lo rechazaba de forma explícita. Lo que hacía que el conductismo pareciera tan radical era la afirmación de que las cuestiones tradicionales no tienen respuestas en los términos en que se plantean tradicionalmente: que el progreso de la psicología exige una revisión a fondo de las suposiciones, el vocabulario, y muy especialmente la ontología de las explicaciones de los procesos mentales basadas en el sentido común.

Si los conductistas hubieran conseguido su objetivo habrían realizado una de las principales revoluciones conceptuales de la historia de la ciencia. No debe sorprendernos que no lo consiguieran. Lo que se proponían hacer era cambiar de arriba abajo la compleja e intrincada estructura de los conceptos mentales que es consecuencia de los intentos realizados durante milenios por llegar a entendernos unos a otros y a nosotros mismos. Quizá se pueda discutir si los resultados de la investigación científica acabarán demostrando que hay que abandonar esta tradición. Pero parece lógico esperar que el proceso se produzca poco a poco —una operación de dentro a fuera—. Sería realmente extraordinario si, al explicar la conducta, pudiéramos arreglárnoslas con categorías explicativas concebidas *de novo* y, como dice John Austin en un contexto semejante, en una tarde libre.

En contraste, la psicología cognitiva contemporánea es, en rasgos generales, conservadora en su actitud hacia la tradición basada en el sentido común. Indudablemente, la flora y la fauna de la psicología han proliferado con gran abundancia, y se postulan sorprendentes procesos mentales a diestro y siniestro. Sin embargo, en el centro del cuadro, el *explicandum* fundamental es el organismo y sus actitudes proposicionales: lo que cree, lo que aprende, lo que quiere y teme, lo que percibe que ocurre. Es decir, los psicólogos cognitivos aceptan lo que los conductistas trataban de rechazar por todos los medios: la *facticidad* de las atribuciones de actitudes proposicionales a los organismos y la necesidad consiguiente de explicar cómo llegan los organismos a tener hacia las proposiciones las actitudes que tienen de hecho.

Lo que hay de *no*-tradicional en el movimiento, si mi reconstrucción es correcta, es la explicación de las actitudes proposicionales que propone: tener una determinada actitud proposicional es estar en una determinada relación con una representación interna. Es decir, por cada una de las actitudes proposicionales (generalmente en número infinito) que el organismo puede tener existe una representación interna y una relación tal que estar en dicha relación con respecto a dicha representación sea nomológicamente necesario y suficiente para (o nomológicamente idéntico a) tener la actitud proposicional. Por ello, lo menos que se le exige a una psicología cognitiva empíricamente adecuada es que especifique, para cada una de las actitudes proposicionales, la representación interna y la relación que, en este sentido, le corresponde. Las actitudes ante las proposiciones son, en ese sentido, «reducidas» a actitudes con respecto a fórmulas, si bien las fórmulas están expresadas en un código interior propio.

Por eso, tener una actitud proposicional es estar en cierta relación con una representación interna. En especial, tener una actitud proposicional es estar en cierta relación *computacional* con una representación interna. Lo que se pretende afirmar con ello es que la secuencia de hechos que determina causalmente el estado mental de un organismo se podría describir como una secuencia de pasos de una derivación si es que hay alguna posibilidad de describirla en el vocabulario de la psicología. Más exactamente: los estados mentales son relaciones entre organismos y representaciones internas, y los estados mentales causalmente interrelacionados se suceden entre sí según unos principios computacionales que se aplican formalmente *a las representaciones*. Este es el sentido en que las representaciones internas constituyen el dominio de los procesos de datos que informan la vida mental. En resumen, un aspecto esencial de las teorías cognitivas es que tratan de interpretar las transformaciones físicas (causales) en cuanto transformaciones de la información, teniendo como resultado el de mostrar la racionalidad de los procesos mentales. De forma algo semejante, la coherencia de un texto aparece cuando una secuencia de formas ortográficas/geométricas se interpreta como secuencia de oraciones de un lenguaje. Si, tal como señala Quine, la traducción es la empresa en que hacemos lo que está de nuestra parte en favor de la racionalidad de los textos, la psicología cognitiva es aquella en que hacemos todo lo que está de nuestra parte en beneficio de la racionalidad de los procesos mentales en general.

Esto es, en mi opinión, un marco de referencia para una ciencia. En cuanto tal, presenta ciertas exigencias sobre el mundo. No será posible elaborar una psicología del tipo que he tratado de esbozar a no ser que los organismos tengan descripciones

pertinentes en cuanto casos concretos de un sistema formal u otro. Naturalmente, lo que hay que tener en cuenta es lo de «pertinente». La pertinencia requiere a) que haya un procedimiento general y plausible para asignar fórmulas del sistema a los estados del organismo, b) que las secuencias causales que determinan las actitudes proposicionales resulten ser derivaciones en el caso de la asignación, c) que por cada actitud proposicional del organismo haya un estado causal del organismo tal que c1) se pueda interpretar el estado como relación con una fórmula del sistema formal y c2) estar en ese estado sea nomológicamente necesario y suficiente para (o contingentemente idéntico con) tener la actitud proposicional.

Es evidente, creo yo, que los puntos a) - c) constituyen constricciones *sustantivas* sobre las teorías psicológicas: no toda asignación de expresiones de un sistema formal a los estados causales de un organismo conseguirá desplegar secuencias de dichos estados en cuanto derivaciones, y esto sigue siendo cierto aun en el caso de que no nos preocupemos demasiado de qué es lo que debe figurar como sistema formal o lo que debe figurar como derivación. Podríamos imaginar, por ejemplo, una asignación de oraciones inglesas a nuestros propios estados fisiológicos de tal manera que, cualquiera que sea el estado nomológicamente necesario y suficiente para creer que va a llover, ese estado estuviera emparejado, por ejemplo, con la oración «ya no hay osos hormigueros». En efecto, según esta atribución, creer que va a llover es estar en una relación determinada con la oración sobre los osos hormigueros. Evidentemente, *esta* atribución de fórmulas a los estados causales no encaja con los puntos a) - c) ya que, en términos generales, las consecuencias *causales* de creer que va a llover no se pueden emparejar de forma coherente con las consecuencias *lógicas* de «ya no hay osos hormigueros». Las relaciones causales entre los estados del organismo no respetan, en ese sentido, las relaciones semánticas entre las oraciones inglesas en la atribución propuesta. Sin embargo, lo que los puntos a) - c) exigen de las teorías psicológicas es precisamente que mantengan esta relación de respeto. Por tanto, si queremos atenernos a los puntos a) - c) sería conveniente que nos cercioráramos cuando menos de que estar en el estado causal que emparejamos con «Ya no hay osos hormigueros» sea, en general, nomológicamente suficiente para estar en el estado que emparejamos con «No hay osos hormigueros» pues, en general, las personas que creen lo que expresa la primera oración creen también lo que expresa la segunda.

Aquí está realmente el punto de partida: ¿Qué clase de sistema formal será lo suficientemente rico como para suministrar el vehículo de la representación interna? ¿Qué clase de operaciones con las fórmulas de este sistema pueden considerarse como operaciones computacionales? ¿Qué secuencias de estas fórmulas constituyen «derivaciones» en el sentido requerido? ¿Qué relaciones entre organismos y fórmulas son tales que el hecho de estar *en* esas relaciones explique el que se tengan actitudes proposicionales, y qué actitudes proposicionales acompañan cada una de las relaciones? ¿Qué principios atribuyen fórmulas a los estados causales? ¿Qué estados causales (y según qué descripciones) son aquellos a los que se atribuyen las fórmulas? es muy poco lo que se sabe sobre la forma de responder a estas preguntas, ni yo he tratado, en la presente obra, de demostrar cómo hay que hacerlo. Mis aspiraciones han sido modestas: El programa dista mucho de estar totalmente aclarado, pero no hay ninguna razón de peso para creer que sea algo esencialmente confuso; el programa trata de problemas que son abstractos cualquiera que sea el criterio aplicado, pero no

hay ninguna razón que nos obligue a negar que se trate de un programa de investigación empírica.

Sin embargo, creo que, de hecho, es bastante claro que el programa no llegará a —no puede— realizarse con la generalidad que aparece en los puntos a) - c). Existen hechos evidentes que parecen constituir un límite para nuestras ambiciones. Quisiera terminar mencionando algunos de ellos.

Los estados mentales, en la medida en que la psicología puede explicarlos, deben ser consecuencias de procesos mentales. Los procesos mentales, según el punto de vista que hemos estado considerando, son procesos en los que se transforman las representaciones internas. Por tanto, los hechos mentales que puede explicar la psicología son aquellos que resultan ser consecuencia de la transformación de las representaciones internas. ¿Cuántos estados mentales entrarían aquí? La argumentación principal de este libro ha sido que viene a ser algo más que ninguno. Pero lo que queremos explicitar ahora es que no se trata de todos. Si esto es cierto, no es posible satisfacer con toda generalidad los puntos a) - c), en especial el b).

En mi opinión, se puede considerar como absolutamente cierto el hecho de que algunas de las actitudes proposicionales que tenemos no son resultado de computaciones. Por supuesto, con ello no queremos decir que no tengan causa ninguna; lo único que decimos es que sus causas no son psicológicas: los hechos que fijan tales estados no tienen interpretación según esa asignación de fórmulas que es la que mejor sirve en conjunto para interpretar la etiología de nuestra actividad mental. Hay ideas que se nos ocurren de repente; a veces nos sorprendemos a nosotros mismos pensando, obsesivamente, en Mónica Vitti; o nos ponemos a dar vueltas sobre si hemos cerrado o no la puerta del sótano. En indudable que en algunas ocasiones dichos estados podrían quedar adecuadamente representados como consecuencias causales de procesos subterráneos de inferencia. Si los freudianos están en lo cierto, eso ocurre más veces de lo que se supone. Pero, indudablemente, no *siempre* ocurre eso. Ciertos estados mentales son, por así decirlo, consecuencia de incursiones violentas del nivel fisiológico; si hubiera que echar la culpa a las ostras que hemos comido, no habría ninguna interpretación *computacional* de la cadena causal que va de ellas hasta la sensación actual de que las cosas, en conjunto, podrían ir mejor.

Como indicaba Davidson (1970), en la vida mental existen lagunas[1]. Las actitudes proposicionales que se fijan mediante computaciones constituyen el objeto material de una ciencia como la que venimos examinando. Pero las demás quedarían fuera, y este hecho permite la posibilidad de que haya fenómenos mentales auténticos que, en principio, no puede explicar una teoría de la cognición.

Quiero destacar este punto porque no hay ninguna razón para creer que las clases de fenómenos mentales que quedan así excluidos del dominio de las teorías del procesamiento de la información se reduzcan a los desechos ocasionales de la vida mental. Por el contrario, algunos de los tipos de hechos mentales más sistemáticos e interesantes quizá sean aquellos sobre cuya etiología cognitiva los psicólogos no tengan absolutamente nada que decir.

---

[1] Lo mismo ocurre en el dominio de cualquiera de las demás ciencias especiales. Si el mundo *es* una secuencia causal continua, puede representarse de esa manera únicamente bajo una descripción física. (Véase lo expuesto en la Introducción.)

El caso más claro es la determinación causal de la sensación. Parece lógico suponer que la integración perceptual del material sensorial se consigue mediante procesos computacionales de carácter general como los mencionados en el Capítulo 1. Pero la etiología del material sensorial debe estar en las interacciones causales entre el organismo y las fuentes de estimulación distal, y estas interacciones, casi por definición, no tienen ninguna representación en el vocabulario psicológico. La psicología cognitiva en cuanto tal no sabe nada sobre el estímulo, excepto lo que se da en una u otra de sus representaciones *proximales*.

Por eso, lo que se *puede* interpretar psicológicamente es cierto de los *efectos* de las interacciones causales entre el organismo y su entorno; a saber, los efectos que forman la base sensorial de la percepción. La etiología de las sensaciones debe ser tratada por una ciencia de otra naturaleza, una ciencia que prediga el estado sensorial del organismo partiendo de las descripciones de las estimulaciones que le afectan. Esto es, evidentemente, lo que ha tratado de realizar la psicofísica clásica. La psicología cognitiva comienza, por así decirlo, donde termina la psicofísica, pero las metodologías de las disciplinas difieren radicalmente. Las verdades psicofísicas expresan la dependencia, según leyes, de hechos *sometidos a descripción psicológica* con respecto a hechos *sometidos a descripción física;* por el contrario las verdades de la psicología cognitiva expresan las dependencias computacionales entre hechos descritos de forma homogénea (psicológicamente). La psicología cognitiva se ocupa de la transformación de representaciones, la psicofísica de la atribución de representaciones a las manifestaciones físicas.

De esta manera, la etiología del material sensorial parece constituir un caso claro en el que la secuencia causal que determina un estado mental no tiene ninguna descripción útil en cuanto secuencia gobernada por reglas de transformaciones de las representaciones. Existen otros casos que son más interesantes, aunque menos claros.

Así, por ejemplo, algunas de las cosas más sorprendentes que hacen las personas —cosas «creativas» como escribir poemas, descubrir leyes, o, genéricamente, tener buenas ideas— no *parecen* especies de procesos gobernados por reglas. Quizá lo sean; quizá haya procedimientos para escribir poemas y la psicología sea cada vez más explícita sobre los mismos según vaya pasando el tiempo. O, lo que quizá resulte más plausible, que haya procedimientos computacionales que dirijan la composición de poemas de acuerdo con *alguna* descripción pero no, por así decirlo, de acuerdo con *esa.* Es decir, puede ocurrir que los procesos que consideremos creativos no formen una clase natural en relación con la explicación psicológica, pero que, sin embargo, cada uno de los *casos* de tales procesos sea un caso de uno u otro tipo de actividad computacional, gobernada por reglas. Las personas que demuestran teoremas y las que hacen soufflés realizan, supongo yo, actividades creativas. De ahí no se desprende que lo que hace el cocinero y lo que hace el matemático sean cosas semejantes según las descripciones que son relevantes respecto a sus explicaciones psicológicas. Las categorías *creativo/aburrido* quizá sólo sean una clasificación cruzada con respecto a la taxonomía que emplea la psicología.

Sin embargo, lo que considero más importante es que el mero hecho de que los procesos mentales creativos sean procesos *mentales* no es garantía de que tengan una explicación en el lenguaje de la psicología bajo *cualquiera* de sus descripciones. Es posible que las buenas ideas (algunas, muchas o todas) sean especies de estados men-

tales que no tienen causas mentales. Como no se sabe absolutamente nada sobre estas cuestiones, no veo ninguna razón para rechazar las intuiciones de las personas creativas sobre las formas en que llegan a actuar creativamente. En este sentido las anécdotas son sorprendentemente coherentes. Las personas que tienen que resolver grandes problemas muchas veces no se ponen a intentar solucionarlos por algún medio intelectual sistemático (o, al menos, si lo hacen, muchas veces no tienen conciencia de que lo están haciendo). Por el contrario, tratan de manipular la situación *causal* con la esperanza de que la manipulación de las causas dará buenos resultados.

Las formas en que las personas hacen esto son notablemente idiosincrásicas. A algunos les da por pasear. Otros se van a la cama. Coleridge y De Quincy fumaban opio. Hardy se iba a ver un partido de cricket. Balzac se ponía el camisón. Proust se sentaba en una habitación encorchada y contemplaba sombreros antiguos. Dios sabe qué haría De Sade. Es posible, ciertamente, que todas estas conductas no sean más que supersticiones. Pero existe el mismo grado de probabilidad de que no lo sean. No hay ninguna razón de principio que excluya la posibilidad de que los estados mentales considerados de gran valor sean a veces resultado de causas (literalmente) no-racionales. Quizá la psicología cognitiva no tenga nada que decir sobre la etiología de tales estados pues, en el mejor de los casos (véase más adelante), de lo que habla es de los estados mentales que tienen causas mentales. Es posible que estemos trabajando en una parcela muy reducida, aunque todavía no sabemos exactamente cuáles son sus límites.

Hasta ahora nos hemos ocupado de los casos en que los estados mentales no dependen (o, en cualquier caso, quizá no dependan) de causas mentales. Lo importante es que la etiología de tales estados cae, por definición, fuera del dominio de los mecanismos de explicación utilizados por la psicología cognitiva; ésta trata de la forma en que se estructura la racionalidad, es decir, de cómo unos estados mentales dependen de otros.

Pero de hecho, la situación puede ser todavía peor. La explicación cognitiva requiere no sólo estados mentales interrelacionados causalmente, sino también estados mentales cuyas relaciones causales respeten las relaciones semánticas que rigen entre las fórmulas dentro del sistema representacional interno. Lo que nos interesa ahora es que puede haber estados mentales cuya etiología escapa a una explicación cognitiva por estar relacionados con sus causas de tal manera que cumplen la primera condición pero no la segunda.

A primera vista parece que estos casos se dan en abundancia, aunque, evidentemente, cualquier explicación sobre la etiología de un estado mental estará a merced de lo que se comprueba empíricamente; lo mejor que podemos hacer es presentar ejemplos plausibles. He aquí uno: alguien desea acordarse, en algún momento de la jornada, de que debe enviar una felicitación a un amigo. Para ello, se cambia el reloj de mano, pues sabe que lo mirará alguna vez y que, cuando lo haga, pensará que tiene que enviar una felicitación. Parece que lo que se da aquí es una conexión causal directa entre dos estados mentales (ver que el reloj está en la otra mano y acordarse de hacer algo), pero un tipo de conexión sobre el que la psicología cognitiva no tiene nada que decir. Aunque los estados mentales estén relacionados causalmente, no lo están en virtud de su *contenido;* compárese con el caso de quien se acuerda de que

debe enviar una felicitación a un amigo cuando a) oye, y b) entiende un ejemplo de elocución del tipo «envía una felicitación a tu amigo».

Creo que es probable que haya muchos tipos de ejemplos de relaciones causales-pero-no-computacionales entre estados mentales. Probablemente, muchos procesos asociativos tienen este carácter, como ocurre quizá con muchos de los efectos de la emoción sobre la percepción y la creencia. Si esta idea es cierta, estamos ante auténticos ejemplos de relaciones causales entre estados mentales que, sin embargo, quedan fuera de los dominios de la explicación psicológica (cognitiva). Lo que *puede* hacer el psicólogo cognitivo es especificar los estados que se relacionan de esta manera y decir *que* lo están. Pero, desde el punto de vista psicológico, la existencia de estas relaciones no pasa de ser una cuestión de mero hecho; su explicación queda en manos de una investigación de nivel inferior (probablemente biológica).

De todas las maneras, no se puede establecer a priori qué aspectos de la vida mental se pueden tratar naturalmente dentro del tipo de marco teórico que se ha considerado en este libro. Dicho tratamiento exige que un estado mental sea analizable en cuanto relación con una representación, y que sus antecedentes causales (o consecuentes o ambas cosas) sean analizables en cuanto relaciones con representaciones relacionadas semánticamente. Esto es, en mi opinión, una condición para que se dé una relación *racional* entre eventos de la vida mental, y supongo que, por definición, sólo las relaciones que son racionales en este sentido amplio pueden tener posibilidad de ser analizadas como computacionales. Pero no todo evento mental tiene una causa mental; a fortiori, no todo evento mental está relacionado racionalmente con causas mentales. El universo del discurso cuya población está formada por los eventos mentales relacionados racionalmente constituye, en una primera aproximación, el dominio natural de una psicología cognitiva. La amplitud de este dominio es algo que queda sometido a la investigación empírica, pero tendría gracia que excluyera gran parte de lo que han estudiado tradicionalmente los psicólogos.

La forma de entender la vida mental que he propuesto puede resultar decepcionante por la modestia de sus aspiraciones. Como he dicho antes, con los instrumentos que tenemos a nuestra disposición lo más que podemos esperar es una teoría de la estructura de la racionalidad; quizá, lo más a que puede aspirar una psicología no reductiva. Aunque a mí me parece que esto es mucho, a muchos les parecerá demasiado poco. A quienes son de esta opinión, mis propuestas serán también decepcionantes en otro sentido.

Un tema muy repetido dentro de este libro ha sido que las operaciones mentales se definen en relación con las representaciones. Puede darse, sin embargo, un estado de ánimo en que esto pueda parecer una trampa. Es fácil representarse la mente como si estuviera obligada a contentarse con unos sucedáneos de la realidad, incapaz, por la naturaleza de las cosas, de entrar en contacto con el mundo exterior. Y es fácil pasar de aquí a un deseo indefinido de inmediatez epistémica; deseo que no es menos apasionado por el hecho de ser en gran parte incoherente. La imposibilidad de decir lo que se quiere es perfectamente compatible con un deseo intenso; de ahí, las fantasías bergsonescas de los gurús de la costa Oeste, como el desaparecido Aldous Huxley.

Por eso, conviene insistir en que esta concepción de la mente *no es* la que yo he desarrollado, ni está implicada por ninguna de mis afirmaciones. En primer lugar,

suponer que los estados mentales son analizables en cuanto relaciones con las representaciones no es eliminar la probabilidad de que sean también analizables en cuanto relaciones con los objetos del mundo. Por el contrario, en la situación epistémicamente normal se entra en relación con una parte del mundo precisamente *a través* de la propia relación con su representación; en una situación normal, si estoy pensando en Mary, es *Mary* en quien estoy pensando. Pensar en Mary es *(inter alia)* representar a *Mary* de una determinada manera; no es, por ejemplo, representar la representación de Mary de esa manera.

No hay, por tanto, ninguna razón en principio para decir que una teoría representacional de la mente deba caer necesariamente en el solipsismo. Además, el tipo de teoría representacional que he estado defendiendo está especialmente imposibilitado para incurrir en esa actitud. Según mi explicación, la secuencia de hechos desde el estímulo a la respuesta es característicamente una secuencia *causal;* en especial, la secuencia de hechos desde el estímulo distal a la representación proximal es característicamente causal. Si este punto de vista es correcto, el solipsismo no puede serlo; no hay efectos producidos por las cosas que no existen.

En cuanto al carácter inmediato, es algo que se puede comprobar con facilidad, aunque no en ninguno de los sentidos que habrían complacido a Huxley. Como nuestros estados epistémicos son consecuencia física de causas físicas, las relaciones epistémicas son inmediatas en cualquiera que sea el sentido en que lo son las relaciones causales, y esto debe ser suficientemente inmediato para cualquiera. Por otra parte, estas relaciones no suelen estar causalmente explicadas en conformidad con las descripciones cuyo cumplimiento las hace epistémicas. Los mismos hechos son epistémicos e inmediatos pero, téngase bien en cuenta, no en los mismos sentidos. No es posible eludir esta situación; no serviría de nada, por ejemplo, dar un cambio y vivir con los animales. Ellos corren la misma suerte.

Nuestras transacciones causales con el mundo son, supongo yo, explicables todas ellas en el vocabulario de la física. Pero las consecuencias epistémicas de dichas transacciones no pueden serlo, pues las propiedades del mundo con que estamos relacionados epistémicamente no son, generalmente, sus propiedades físicas. Considero que esto es una cuestión de hecho, sin más. Podríamos imaginarnos algún organismo que sólo tuviera conocimiento de aquellos rasgos del mundo que deben tenerse en cuenta en las explicaciones causales de *lo que* sabe el organismo. Pero, de hecho, no existen tales organismos. En realidad, es el hecho de que no los haya lo que confirma en último término el principio metodológico expuesto en la Introducción: el vocabulario teórico de la psicología es muy diferente del vocabulario teórico de la física. Llegamos así, tras un largo recorrido, a un punto en que se funden inextricablemente los supuestos metodológicos y empíricos de la investigación. Parece un buen lugar para detenerse.

BIBLIOGRAFIA

Atherton, M. y Schwartz, R. (1974), «Linguistic Innateness and Its Evidence». *Journal of Philosophy,* LXXI, no. 6, 155-168.

Bar-Hillel, Y. (1970), *Aspects of Language,* The Magnes Press, The Hebrew University, Jerusalén.

Bartlett, F. C. (1961), *Remembering,* (primera edición, 1932). Cambridge Univ. Press, Londres y Nueva York.

Berlyne, D. E. (1965), *Structure and Direction in Thinking,* Wiley, Nueva York.

Bever, T. G. (1970), «The cognitive basis for linguistic structures». *En Cognition and the Development of Language.* (J. R. Hayes, ed.), Wiley, Nueva York.

Block, N. J. and Fodor, J. (1972), «What psychological states are not». *Philosophical Review,* 81, 159-181.

Blumenthal, A. L. (1966), «Observations with self-embedded sentences». *Psychonomic Science,* 6, 453-454.

Bransford, J. D. y Franks, J. J. (1971), «The abstraction of linguistic ideas». *Cognitive Psychology,* 2, 331-350.

Brewer, W. F., «There is no convincing evidence for operant or classical conditioning in adult humans». *Cognition and Symbolic Processes.* (W. B. Weiner and D. S. Palermo, eds.), de próxima publicación.

Broadbent, D. E. (1958), *Perception and Communication,* Pergamon, Oxford.

Brooks, L. R. (1968), «Spatial and verbal components of the act of recall». *Canadian Journal of Psychology,* 22, 349-368.

Brown, R. (1958), *Words and Things,* The Free Press, Nueva York.

Brown, R. (1970), «How shall a think be called?» en *Psycholinguistics,* The Free Press, Nueva York.

Bruner, J. S. (1957), «On perceptual readiness». *Psychological Review,* 64, 123-152.

Bruner, J. S., Goodnow, J. J. y Austin, G. A. (1956), «A Study of Thinking», Wiley, Nueva York. (Paperback Wiley Science Editions, 1962.)

Bruner, J. S., Olver, R. R. y Greenfield, P. M. (1966), *Studies in Cognitive Growth,* Wiley, Nueva York.

Bryant, P. E. (1974). *Perception and Understanding in Young Children,* Basic Books, Nueva York.
Bryant, P. E. y Trabasso, T. (1971). «Transitive inferences and meaning in young children». *Nature,* 232, 456-458.
Capranica, R. R. (1965). *The Evoked Vocal Response of the Bullfrog: A Study of Communication by Sound,* MIT Press, Cambridge, Massachusetts.
Carnap, R. (1956), *Meaning and Necessity,* Univ. of Chicago Press, Chicago, Illinois.
Chihara, C. y Fodor, J. (1965), «Operationalism and ordinary language». *American Philosophical Quarterly.* 2(4), 281-295.
Chomsky, N. (1957), *Syntactic Structures,* Mouton & Co., The Hague.
Chomsky, N. (1959), Review of Skinner's «Verbal Behavior», *Language,* 35, 26-58.
Chomsky, N. (1965), *Aspects of the Theory of Syntax.* MIT Press, Cambridge, Massachusetts.
Chomsky, N. (1969), «Linguistics and Philosophy». En *Language and Philosophy.* (S. Hook, ed.), N. Y. Univ. Press, Nueva York.
Clark, H. H. y Chase, W. G. (1972), «On the process of comparing sentences against pictures». *Cognitive Psychology,* 3, 472-517.
Clifton, C. y Odom, P. (1966), «Similarity relations among certain English sentence constructions». *Psychological Monographs,* 80 (Whole No. 613).
Collins, A. M. y Quillian, M. R. (1969), «Retrieval time from semantic memory». *Journal of Verbal Learning and Verbal Behavior,* 8, 240-247.
Cooper, L. A. y Shepard, R. N. (1973), «Chronometric studies of the rotation of mental images». *Visual Information Processig* (W. G. Chase, ed.), Academic Press, Nueva York.
Crowder, R. G. y Morton, J. (1969), «Precategorical acoustic storage» (PAS). *Perception and Psychophysics,* 5, 365-371.
Davidson, D. (1967), «Truth and meaning». *Synthese,* 17, 304-323.
Davidson, D. (1970), «Mental events». En *Experience and Theory,* (L. Forster y J. Swanson, eds.), Univ. of Massachusetts Press, Amherst, Massachusetts.
Dennett, D. C. (1969), *Content and Consciousness,* Humanities Press, Nueva York.
Dennett, D. C. (1972), *Skinner Skinned.* Sin publicar.
Dreyfus, H. L. (1972), *What Computers Can't Do: A Critique of Artificial Reason.* Harper, Nueva York.
Dulany, D. E., Jr. (1968), «Awareness, rules, and propositional control: a confrontation with S-R behavior theory». En *Verbal Behavior and General Behavior Theory* (T. R. Dixon and D. L. Horton, eds.), Prentice-Hall, Englewood Cliffs, New Jersey.
Fillmore, C. (1971), «Entailment rules in a semantic theory». En *The Philosophy of Language* (J. Rosenberg y C. Travis, eds.), Prentice Hall, Englewood Cliffs, New Jersey.
Fodor, J. A. (1968), *Psychological Explanation,* Random House (Smithsonian Inst. Press), Nueva York.
Fodor, J. A. (1970), «Three reasons for not deriving "kill" from "cause to die"». *Linguistic Inquiry,* 1, 429-438.
Fodor, J. A. (1972), «Some reflections on L. S. Vygotsky's "Thought and Language"», *Cognition,* 1(1), 83-95.
Fodor, J. A. «Special sciences». *Synthese,* de próxima aparición.
Fodor, J. A., Bever, T. y Garrett, M. (1974), *The Psychology of Language: An Introduction to Psychgolinguistics and Generative Grammar,* McGraw-Hill, Nueva York.
Fodor, J. A., Fodor, J. D. y Garrett, M. «The psychological reality of semantic representations». *Linguistic Inquiry,* de próxima aparición.
Fodor, J. A., Garrett, M. y Bever, T. (1968), «Some syntactic determinants of sentential complexity, II: Verb Structure». *Perception and Psychophysics,* 3, 453-461.
Fodor, J. A., Garrett, M. y Brill, S. L. (1975), «Pe, ka, pu: the perception of speech sounds

in prelinguistic infants». *M.I.T. Quarterly Progress Report,* Enero, 1975. (Resumido en *Science*).
Fodor, J. D. *Semantics,* de próxima aparición.
Forster, K. I. y Olbrei, I. (1973), «Semantic heuristics and semantic analysis». *Cognition,* 2, 319-348.
Garrett, M. y Fodor, J. A. (1968), «Psychological theories and linguistic constructs». En *Verbal Behavior and General Behavior Theory* (T. R. Dixon y D. L. Horton, eds.), Prentice-Hall, Englewood Cliffs, New Jersey.
Gibson, J. J. (1966), *The Senses Considered as Perceptual Systems,* Houghton, Boston, Massachusetts.
Goodman, N. (1965), *Fact, Fiction and Forecast,* Bobbs-Merrill, Indianápolis, Indiana.
Goodman, N. (1968), *Languages of Art,* Bobbs-Merrill, Indianápolis, Indiana.
Greene, J. (1972), *Psycholinguistics: Chomsky and Psychology,* Penguin, Harmondsworth, Middlesex, Inglaterra.
Gregory, R. L. (1966), *Eye and Brain: The Psychology of Seeing,* McGraw-Hill, New York.
Grice, H. P. (1957), «Meaning». *The Philosophical Review,* LXVI. (Reimpreso en *Problems in the Philosophy of Language* (T. Olshewsky, ed.), Holt, Nueva York.
Haber, R. N. (1966), «Nature of the effect of set on perception». *Psychological Review,* 73, 335-351.
Halle, M. y Stevens, K. N. (1964), «Speech recognition: a model and a program for research». En *The Structure of Language: Readings in the Philosophy of Language* (J. A. Fodor y J. J. Katz, eds.), Prentice-Hall, Englewood Cliffs, New Jersey.
Harman, G. (1969), «Linguistic competence and empiricism». En *Language and Philosophy* (S. Hook, ed.), N. Y. Univ. Press, Nueva York.
Heider, E. (1971), «Natural categories». *Proceedings, 79th Annual Convention, American Psychological Association.*
Helke, M. (1971), «The grammar of English reflexives». M.I.T. Tesis doctoral.
Hull, C. L. (1943), *Principles of Behavior,* Appleton, Nueva York.
Hume, D. (1960), *A Treatise of Human Nature,* Vol. 1. (Primera edición 1739), Dent, Londres.
Jarvella, R. J. (1970), *Effects of syntax on running memory span for connected discourse.* Psychonomic Science, 19, 235-236.
Johnson-Laird, P. N. y Stevenson, R. (1970). «Memory for syntax». *Nature,* 227, 412-413.
Julesz, B. (1965), «Texture and visual perception». *Scientific American,* febrero 38-48.
Kant, I. (1953), *Critique of Pure Reason,* N. K. Smith (Trad.). Macmillan, Nueva York. (Primera edición).
Katz, J. J. (1966), *The Philosophy of Language,* Harper, Nueva York.
Katz, J. J. (1972), *Semantic Theory,* Harper, Nueva York.
Katz, J. J. y Fodor, J. A. (1963), «The structure of a semantic theory». *Language,* 39, 170-210.
Katz, J. J. y Postal, P. M. (1964), *An Integrated Theory of Linguistic Descriptions,* MIT Press, Cambridge, Massachusetts.
Kripke, S. (1972), «Naming and necessity». En *Semantics of Natural Language* (G. Harmon y D. Davidson, eds.), Humanities Press, Nueva York.
Lackner, J. R. y Garrett, M. (1973), «Resolving ambiguity: effects of biasing context in the unattended ear». *Cognition,* 1, 359-372.
Lakoff, G. (1970a), *Irregularity in Syntax,* Holt, Nueva York.
Lakoff, G. (1970b), «Linguistics and natural logic. *Studies in Generative Semantics No.* 1. Trabajos de fonética, Univ. of Michigan, Ann Arbor, Michigan.
Letvin, J., Maturana, H., Pitts, V. y McCulloch (1961), «Two remarks on the visual system of the frog». En *Sensory Communication* (W. Rosenblith, ed.), MIT Press, Cambridge, Massachusetts.

Lewis, D. K. (1970), «General semantics». *Synthese,* 22, 18-67.

Liberman, A., Cooper, F. S., Shankweiler, D. P. y Studdert-Kennedy, M. (1967), «Perception of the speech code». *Psychological Review,* 74, 431-461.

Loewenstein, W. R. (1960), «Biological transducers». *Scientific American,* agosto. (También en *Perception: Mechanisms and Models, Readings from Scientific American* (1972), Freeman, San Francisco, California.

Luria, A. R. (1968), *The Mind of the Mnemonist,* Basic Books, Nueva York.

Malcolm, N. (1962), *Dreaming,* Humanities Press, Nueva York.

Marslen-Wilson, W. (1973), «Speech shadowing and speech perception». Tesis doctoral. Massachusetts Inst. of Technology, Cambridge, Massachusetts.

McCawley, J. D. (1970), «Syntactic and logical arguments for semantic studies». *Transcript at the Fifth International Seminar of Theoretical Linguistics (sesión 2),* Tokyo, sin publicar.

McCawley, J. D. (1971), «Prelexical syntax». En *Georgetown Monograph Series on Languages and Linguistics,* (R. J. O'Brien, ed.).

Mehler, J. (1963), «Some effects of grammatical transformation on the recall of English sentences». *Journal of Verbal Learning and Verbal Behavior,* 2, 346-351.

Miller, G. A., Galanter, E. y Pribram, K. H. (1960), *Plans and the Structure of Behavior,* Holt, Nueva York.

Miller, G. A. y Johnson-Laird, P. N. *Perception and Language,* de próxima publicación.

Neiser, U. (1967), *Cognitive Psychology,* Appleton, Nueva York.

Newell, A. y Simon, H. A. (1972), *Human Problem Solving,* Prentice-Hall, Englewood Cliffs, New Jersey.

Norman, D. A. (1969), *Memory and Attention,* Wiley, Nueva York.

Oppenheim, P. y Putnam, H. (1958), «Unity of science as a working hypothesis». En *Minnesota Studies in the Philosophy of Science,* Vol. II, (H. Feigl, M. Scriven y G. Mazwell, eds.), Univ. of Minnesota Press, Minneapolis, Minnesota.

Osgood, C. E. (1957), «Motivational dynamics of language behavior». *Nebraska Symposium on Motivation,* 5, 348-424.

Paivio, A. (1971). *Imagery and Verbal Processes,* Holt, Nueva York.

Penfield, W. y Roberts, L. (1959), *Speech and Brain Mechanisms,* Princeton Univ. Princeton, New Jersey.

Perky, C. W. (1910), «An experimental study of imagination». *American Journal of Psychology,* 21, 422-452.

Piaget, J. (1954), *The Construction of Reality in the Child,* Basic, Nueva York.

Piaget, J. (1970), *Structuralism,* Basic, Nueva York.

Posner, M. I., Boies, S. J., Eichelman, W. H. y Taylor, R. L. (1969), «Retention of visual and name codes of single letters». *Journal of Experimental Psychology Monograph,* 79 (1, P. 2).

Pribram, K. H. (1971), *Languages of the Brain,* Prentice-Hall, Englewood Cliffs, New Jcrsey.

Putnam, H. (1960a), «Dreaming and depth grammar». En *Analytic Philosophy,* (R. J. Butler, ed.), Barnes & Noble, Nueva York.

Putnam, H. (1960b), «Minds y machines». En *Dimensions of Mind* (S. Hook, ed.), N. Y. Univ. Press, Nueva York.

Putnam, H. *The meaning of meaning.* (De próxima publicación).

Ratliff, F. (1961), «Inhibitory interaction and the detection and enhancement of contours». En *Sensory Communication* (W. Rosenblith, ed.), MIT Press, Cambridge, Massachusetts.

Richardson, A. (1969), *Mental Imagery,* Springer, Nueva York.

Rosenberg, S. (1974), «Modelling semantic memory: effects of presenting semantic information in different modalities». (Tesis doctoral), Carnegie-Mellon University.

Ross, J. T. (1967), «Constraints on variables in syntax» (Ph. Tesis doctoral), Massachusetts

Inst. of Technology, Cambridge, Massachusetts.
Russell, B. (1905), «On denoting». *Mind,* XIV, 479-493.
Ryle, G. (1949), *The Concept of Mind,* Barnes & Noble, Nueva York.
Sachs, J. S. (1967), «Recognition memory for syntactic and semantic aspects of connected discourse». *Perception and psychophysics,* 2, 437-442.
Savin, H. B. (1973), «Meanings and concepts: a review of Katz's «Semantic Analysis», *Cognition,* 2(2), 212-238.
Schlesinger, I. M. (1966), «Sentence structure and the reading process». (Ph. Tesis doctoral), The Hebrew University, Jerusalén.
Segal, S. J. y Gordon, P. (1968), «The Perky effect revisited: paradoxical thresholds or signal detection error?» Informe presentado en la 39 Reunión Anual de la Eastern Psychological Association.
Skinner, B. (1957), *Verbal Behavior,* Appleton, Nueva York.
Skinner, B. F. (1971), *Beyond Freedom and Dignity.* Knopf, Nueva York.
Slobin, D. I. (1966), «Grammatical transformations and sentence comprehension in childhood and adulthood». *Journal of Verbal Learning and Verbal Behavior,* 5, 219-227.
Sperry, R. W. (1956), «The eye and the brain». *Scientific American,* Mayo, 48-52.
Stromeyer, C. F. y Psotka, J. (1970), «The detailed texture of eidetic images». *Nature,* 225, 346-349.
Sutherland, I. E. (1970), «Computer displays», *Scientific American,* 222, No. 6, 56-81.
Sutherland, N. S. (1960), «Theories of shape discrimination in octopus». *Nature, London,* 186, 840-844.
Teuber, H. L. (1960), «Perception». En *Handbook of Physiology,* Vol. 3 (J. Field, H. W. Magoun y V. E. Hall, eds.), Amer. Phys. Soc., Washington, D. C.
Thorpe, W. H. (1963), *Learning and Instinct in Animals,* Methuen, Londres.
Tolman, E. C. (1932), *Purposive Behavior in Animals and Men,* Century, Nueva York.
Triesman, A. (1964), «Verbal cues, language and meaning in attention». *American Journal of Psychology,* 77, 206-214.
Vendler, Z. (1972), *Res Cogitans,* Cornell Univ. Press, Ithaca, Nueva York.
Vygotsky, L. S. (1965), *Thought and Language,* MIT Press, Cambridge, Massachusetts.
Walker, E., Gough, P. y Wall, R. (1968), Grammatical relations and the search of sentences in immediate memory». *Proceedings of the Midwestern Psychological Association,* 1968.
Wallach, L. (1969), «On the basis of conservation». En *Studies in Cognitive Development* (D. Elkind y J. Flavell, eds.), Oxford Univ. Press, Londres y Nueva York.
Wanner, E. (1968), «On remembering, forgetting, and understanding sentences: a study of the deep structure hypothesis». (Ph. Tesis doctoral), Harvard University. Mouton, The Hague: de próxima aparición.
Wason, P. C. y Johnson-Laird, P. N. (1972), *Psychology of Reasoning: Structure and Content,* Batsford, Londres, Harvard University Press, Cambridge, Massachusetts.
Werner, H. y Kaplan, B. (1963), *Symbol Formation: An Organismic-Developmental Approach to the Psychology of Language and Expression of Thought,* Wiley, Nueva York.
Whorf, B. L. (1956), *Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf* (John B. Carroll, ed.), Wiley, Nueva York.
Wittgenstein, L. (1953), *Philosophical Investigations,* Blackwell, Oxford.
Young, R. K. (1968), «Serial learning». En *Verbal Behavior and General Behavior Theory* (T. Dixon y D. Horton, eds.), Prentice-Hall, Englewood Cliffs, New Jersey.

INDICE ANALITICO

Ahterton, M., 83.
atención, 175-179.
Agustín, 82.
Austin, G., 55.
Austin, J. L., 213.

Balzac, H., 218.
Bartlett, F. G., 93, 188.
conductismo, 24-30, 52-53.
Berkeley, G., 209.
Berlyne, D. E., 189.
Bever, T. G., 184, 186-187.
Boies, S. J., 202.
Bransford, J. D., 93.
Brewer, W. F., 56, 211.
leyes puente, 32-33, 41.
Brill, S., 75.
Broadbent, D. E., 175, 177, 179.
Brooks, L. R., 201.
Brown, R., 167, 193, 208.
Bryant, P. E., 100, 109.

Capranica, R. R., 66.
Carnap, R., 164.
Cartesianismo, 25, 27.
Cassirer, E., 69.
oraciones incrustadas centrales, 185.
Chase, W. G., 127, 162.
Chihara, C., 124.
decisión, teorías de, 48-53.
Chomsky, N., 76, 116, 117, 125, 157.
Clark, H., 127, 162.
hipótesis de codificación, 130.
Coleridge, S., 218.
Collins, A. M., 165.
comunicación, 118-123.
compiladores, 131-134.
ordenadores, 83-84, 90-91.
aprendizaje de conceptos, teorías del 54-61.
Cooper, L. A., 202.
Crowder, R. G., 177.

Darwin, C., 76.
Davidson, D., 37, 216.
definición, 140-169.
  en la semántica generativa, 144-163.
  en la semántica interpretativa, 139-142.
  por el uso, 142.
Dennett, D., 91, 204, 205, 206, 211.
de Quincy, 218.
de Sade, 218.
Dewey, J., 107.
Dreyfus, B., 80.
Dulaney, D. E., 211.

Eichelman, W. H., 202.
equi NP-deletion (supresión por equivalencia de SNs), 147-148.

Fillmore, 164.

Fodor, J. A., 25, 67, 75-76, 93, 102, 116, 118, 124, 125, 128, 130, 139, 144, 164, 180, 184.
Fodor, J. D., 141, 163, 164.
Forster, K. I., 186.
Franks, J. U., 93.
Freud, S., 71.

Garrett, M. F., 75, 118, 177-179, 184.
Gaugh, P., 185.
Gibson, J., 67-68.
Galanter, E., 162.
Goodman, N., 59.
Goodnow, J. C., 55.
Gordon, P., 203.
Greene, J., 117.
Gregory, R., 69.
Grice, H. P., 119.

Haber, R. N., 202-203.
Hardy, G. H., 218.
Halle, M., 210.
Heider, E., 112, 168, 188.
Helke, M., 155.
Hull, C. L., 189.
Hume, D., 209.

imágenes, 190-211.
innatismo, 76, 99-113.

Jarvella, R. J., 177.
Johnson-Laird, P., 60, 131.

Kant, I., 203.
Katz, J. J., 76, 135, 141, 163.
Kripki, S., 112.

Lackner, J. R., 177-179.
Lakoff, C., 144, 164.
aprendizaje del lenguaje, 76, 96-113.
Lashley, K., 38.
Letvin, J., 66.
lexicalización, 146.
Liberman, A., 131.
relativismo lingüístico, 102.
Locke, J., 111.
Loewenstein, W. R., 66.
Luria, A. R., 188.

McCawley, J. D., 144, 152-163.
MacCorquodale, K., 117.
Marslen-Wilson, W., 166.
postulados de significado, 163-166.
teoría mediacional del aprendizaje, 189.
Mehler, J., 180.
mensajes, 126-131.
Miller, G. A., 131, 162.
Morton, J., 177.

clases naturales, 34-36, 42-46.
Neisser, U., 101, 210.
Norman, D. A., 133, 162, 188.

Olbrei, I., 186.
textura abierta, 80.
Oppenheim, P., 37.
Osgood, C. E., 189.

Paivio, A., 162, 168, 188, 191, 201, 203.
oraciones pasivas, 185.
Penfield, W., 210.
percepción, teorías de la, 62-70.
Perky, C. W., 203.
Piaget, J., 103-110, 192.
Posner, M. I., 202, 210.
Postal, P. M., 135.
subida del predicado, 145.
Pribram, K. H., 162, 210.
argumento del lenguaje privado, 86-90, 95.
actitudes proposicionales, 89-95, 214-219.
Proust, M., 210, 218.
Psotka, J., 202.
explicación psicológica, 91-93.
psicofísica, 217.
Putnam, H., 37, 78, 112, 124, 168.

Quillian, M. R., 165.
Quine, W. V., 214.

Ratliff, F., 66.
oraciones reflexivas, 149-150, 152-160.
Richardson, A., 201.
Roberts, L., 210.
Rosenberg, S., 128.
Ross, J. T., 151, 160.
Russell, B., 28, 142.
Ryle, G., 24-27.

Savin, H. B., 111.
Schlesinger, I. M., 185.
Schwartz, R., 83.

Segal, S. J., 203.
nivel semántico (de una gramática), 139.
mecanismos sensoriales, 65-66.
Shephard, R. N., 202.
single node constraint (SNC) [constricción del nodo único], 144-146.
Skinner, B. J., 55, 117.
Slobin, D. I., 185.
identidad borrosa, 154.
Sperry, R. W., 128.
etapas (del desarrollo cognitivo), 103-110.
Stevens, K. N., 210.
argumento del «tigre listado», 204-207.
Stromeyer, C., 202.
descripciones estructurales, 126.
Sutherland, I. E., 209.
Sutherland, N. S., 76.

Taylor, R. L., 202.
Teuber, H. L., 69.
Thomson, J. J., 148.
Thorpe, W. H., 54, 109.
fisicismo de instancias o casos, 34, 37-40.
Tolman, E. C., 56.
Trabasso, T., 109.
«traducción», teorías del significado, 134-137.
Treisman, A., 176-179.
Treisman, M., 162.
reglas de verdad, 77-82, 96-99.
fisicismo de tipos o propiedades 34, 38-40.

unidad de la ciencia, 31, 39-40, 43-45.

Vendler, Z., 76.
conducta verbal, causalidad de, 116-118.
Vygotsky, L. S., 55, 103, 167, 193.

Walker, E., 185.
Wall, R., 185.
Wallach, I., 104.
Wason, P. C., 60.
Werner, H., 191.
Whorf, B. L., 99-102.
Wittgenstein, L., 24, 80, 82, 86-90, 99, 196.

Young, R. K., 54.